APPLICATION OF CLUSTERING TECHNIQUES IN DEFINING LEVEL OF SERVICE CRITERIA OF URBAN STREETS

AMIT KUMAR DAS

DEPARTMENT OF CIVIL ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

ROURKELA-769008

2013

APPLICATION OF CLUSTERING TECHNIQUES IN DEFINING LEVEL OF SERVICE CRITERIA OF URBAN STREETS

Thesis

Submitted in partial fulfillment of the requirements
For the degree of

Master of Technology
in
Transportation Engineering

By
Amit Kumar Das
Roll No**. 211CE3240**

Under the guidance of
Prof. P. K. Bhuyan

DEPARTMENT OF CIVIL ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

ROURKELA-769008

2013

NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA-769008

CERTIFICATE

This is to certify that project entitled, "**Application of Clustering Techniques in Defining Level of Service of Urban Streets**" submitted by **AMIT KUMAR DAS** in partial fulfillment of the requirements for the award of **Master of Technology** Degree in **Civil Engineering** with specialization in **Transportation Engineering** at National Institute of Technology, Rourkela is an authentic work carried out by him under my supervision and guidance. To the best of my knowledge, the matter embodied in this Project review report has not been submitted to any other University/ institute for award of any Degree or Diploma.

Prof . P. K. Bhuyan
Department of Civil Engineering
National Institute of Technology,
Rourkela- 769008

ACKNOWLEDGEMENT

I would like to thank several individuals who in one way or another contributed and extended their help in preparation and completion of this study. My sincere thanks to **Dr. P. K. Bhuyan** whose motivation and guidance has been my inspiration in the completion of this research work.

My utmost gratitude to **Prof. M. Panda** former HOD of Civil Engineering Department, NIT Rourkela for providing necessary advice and co-operation throughout my M. Tech study.

I extend my thankfulness to **Dr. S. Sarangi**, Director, NIT Rourkela and **Dr. N. Roy**, HOD, Civil Engineering Department, NIT Rourkela for providing necessary facilities for this research work.

My sincere thanks to all my friends at NIT, Rourkela for making my stay in the campus a pleasant one. The cooperation shown by them is worth noting.

Lastly I would thank my parents and the almighty God for giving me support and courage throughout this study.

Amit Kumar Das
Roll No: 211CE3240

CONTENTS

Abstract

The speed ranges for Level of Service (LOS) categories are not well defined for highly heterogeneous traffic flow on urban streets of India. The LOS analysis procedure followed in India is that developed by HCM 2000. The LOS categories for various urban street classes defined by HCM are apposite for developed countries having homogeneous type of traffic flow. For developing countries like India where the traffic flow is highly heteregenous, LOS should be defined correctly taking into account the traffic and geometric characteristics. In this study an attempt has been made to define the LOS criteria of urban streets. Mumbai the business capital of India was chosen as the study area comprising of 100 street segments on four north-south and one east-west corridor. The total length of these five corridors is 140km. LOS analysis is considered important as this affects planning, design, operational aspects of transportation projects and allocation of limited resources among various competing projects. Second-wise speed data collected using Global Positioning System (GPS) receiver fitted on mobile vehicles was used for this study. Free-flow speed (FFS) data, average travel speeds during both peak and off peak hours and inventory details were collected and used in this study. These data are obtained from secondary source for this research work. Defining level of service is basically classification problems. Cluster analysis is found to be the most suitable technique for solving these classification problems. Four clustering methods namely Clustering Large Applications (CLARA), Self Organizing Tree Algorithm (SOTA), Hard Competitive Learning (hardcl) and Neural gas (ngas) were used to define LOS criteria in this study. Calinski-Harabasz Index, Homogenity Index, Stability Index, Connectivity Index, Average proportion of non-overlap Index, Average distance Index, Average distance between means Insex, Figure of merit Index, PtBiserial Index, Tau Index, GPlus Index, Ratkowsky Index, Duda Index, McClain Index are used in deriving optimum number of clusters.

CLARA, SOTA, hardcl and ngas algorithms were applied in two stages. In the first stage, clustering methods were applied on FFS data and free-flow speeds were classified into four groups corresponding to urban street classes I to IV. In the second phase, clustering methods were applied on average travel speeds on street segments of each class of urban street during both peak and off peak hours. These average travel speeds were classified into six categories for

six levels of service. Results obtained by applying these four methods show speed ranges for urban street classes and level of service categories which are significantly different from the corresponding values mentioned in HCM 2000. Based on the result it was found that vehicles traveled more frequently at poor level of service i.e. LOS "D", "E", "F" than good level of service i.e. LOS "A", "B", "C". The average travel speed of LOS categories are expressed in terms of percentage of FFS of the street classes. The result closely follows the percentage shown in HCM 2010.

Keywords: Level of Service, Global Positioning System, Urban streets, CLARA Clustering, SOTA Clustering, hardcl Clustering, ngas Clustering.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General

The rapid growth of India's urban population has put enormous strains to the developing country. Pertaining to the global trend of economic growth India has to go under rapid urbanization in the last century which has become more pronounced after independence. According to a study conducted by Government of India the urban population of India was recorded to be 286million in the year 2001 which was 27.85percent of the total population. By the year 2011 this population increased to 377million which was 31.16percent of the total population. By 2025 this population is expected to be 600million. The number of towns has increased from 5161 in 2001 to 7935 in 2011. Even at this relatively low level of urbanization (31%), India has the second largest urban population in the world. According to the World Urbanization Prospects (the 1996 Revision), the urban population in the year 2025 will rise to 42.5 per cent which is 566 million (UN Department of Economic and Social Affairs, 1996). In 2002, 58.8 million vehicles were plying on Indian roads. As per Ministry of Road Transport & Highways, Government of India, the annual rate of growth of motor vehicle population in India has been about 10 percent during the last decade. The matter of concern which drags attention is the concentration of the vehicles in few selected cities rather than their total numbers in the country. It is interesting to mention that 32percent of these vehicles ply in metropolitan cities alone.

In developing countries like India the traffic in urban street is composed of several types of vehicles ranging from cars with high speeds to low speed non-motorized vehicles (heterogonous traffic condition). Presently there are no proper methodologies to evaluate Level of Service (LOS) provided by urban streets in India. It is a bare necessity to develop suitable methodologies for level of service analysis procedure for urban streets as these methodologies affect the planning, design, and operational aspects of transportation projects as well as the allocation of limited financial resources among competing transportation projects. This brings in the

importance of suitable methods that should be chosen while defining level of service criteria of urban streets in Indian context.

Urban street level of service is based on average through-vehicle travel speed for the segment, section, or entire street under consideration. The most common technique used to obtain speed data is Floating car method. A passenger records the elapsed time information at predefined check points while the driver drives the car. The recording of elapsed time can be done by pen and paper, audio recorder or with a small data recording device. The advantage of this method is it requires very low skilled technician and the cost involved is very low. The disadvantage of this method is it is a labor intensive method. Labor intensive works involves human errors, often includes both recording errors in the field and transcription errors as the data is put into an electronic format (Turner, *et. al.*,1998). Advancement of computers brought the Distance Measuring Instrument (DMI) as a solution for floating car method, DMI measures the speed distance using pulses from a sensor attached to the test vehicle's transmission (Quiroga and Bullock, 1998). The shortcoming with this method is very complicated wiring is required to install DMI unit to a vehicle. Frequent calibration and verification factors unrelated to the unit are necessary to store making the data file large and which leads to the data storage problem (Turner, et. al. 1998).

With technological advancement Global Positioning System (GPS) has made it easy to maintain large database with great accuracy. The shortcomings of floating car method and DMI for field data collection are replaced with GPS. A GPS receiver is mounted on a vehicle which records the location as latitude and longitude, travel time and travel speed. These tasks can be performed by a single technician with accuracy. The field data collected through GPS receiver should refer to a geographical position. With the use of Geographical Information System (GIS) this task becomes easy. The parameters received through GPS are assigned to the existing geographical data base. Making it easy to understand, defining level of service criteria is a classification problem and clustering methods can be used for its solution. The parameters received through GPS are assigned to the existing geographical data base contained in GIS. This helps in collecting large amount of travel time and speed data with greater convenience, consistency, finer precession and accuracy than the conventional procedure.

The concept of Level of Service (LOS) was first introduced in the 1965 Highway Capacity Manual (TRB, 1965). In the 1965 Highway Capacity Manual (HCM), the level of service was described by six classes from "A" to "F" defined, based on the combination of travel time and the ratio of traffic flow rate to the capacity, because travel time was considered as a dominant factor of the service quality. However these classes were not defined in a quantitative manner. Therefore the concept was again redefined considering several traffic conditions in the HCM version 1985 (TRB, 1985). The measures of levels of service adopted in TRB (1985), which describe the characteristics of traffic conditions under operation, include travel speed, traffic flow rate, and traffic density, for each type of road. The definition of LOS has changed with due course of time and what we are following now is the LOS defined in HCM 2000. Highway Capacity Manual (HCM, 2000) defines LOS as "quality measure describing operational conditions within a traffic stream, generally in terms of such service measures as speed and travel time freedom to maneuver, traffic interruptions, comfort and convenience. " HCM (2000) also designates six levels of service from "A" to "F" for each type of facility, with LOS "A" representing the best operating conditions and LOS "F" the worst.

Several studies have been performed for the derivation of LOS criteria. Cameron (1996) and Baumgaertner (1996) proposed extended criteria from "A" to "J" and "A" to "I" respectively, instead of LOS "A" through "F" as mentioned in HCM 2000, taking into account urban traffic congestion. HCM (2000), IRC (1990) guidelines states that factors like heterogeneity of traffic, speed regulations, frequency of intersections, presence of bus-stops, on-street parking, roadside commercial activities, pedestrian volumes etc. affects the level of service strongly on urban roads. Marwah and Singh (2000) based on their simulation results of benchmark roads and traffic composition, tried to provide classification of level of service categories for urban streets. They classified level of service into four groups (LOS I-IV). Maitra et al. (1999) conducted a study for heterogeneous traffic flow condition on urban roads and redefined the LOS boundaries by quantifying congestion as measure of effectiveness. LOS boundaries have been categorized into nine groups "A" to "I" in the stable zone and one LOS "J" for unstable zone based on quantified congestion.

Defining LOS is a classification problem and can be solved using various clustering techniques. Prassas et al. (1996) applied the cluster analysis tools to a set of traffic engineering data on which deterministic modeling and regression analysis have been applied before. From this study the authors have concluded that cluster analysis is a powerful exploratory technique and helps in identifying several distinct modalities within the traffic data. Cheol and Stephen (2002) used *k*-means, fuzzy and Self Organizing Map (SOM) clustering in a real-time signalized intersection surveillance system in the determination of LOS categories. Wi et al. (1996) explored the feasibility of vehicle classification in real world situations using an integrated model that they have developed from back-propagation Artificial Neural Network (ANN) model and image processors. Yang and Qiao (1998) have developed a classification method for traffic flow states for Chinese highways using neural network approach.

1.2 Statement of the Problem

Urbanization brings with it several consequences-both adverse and beneficial. Rapid urbanization in today's world has triggered several negative effects on the urban road networks like decreasing speed, increasing congestion, increased travel time, decreased level of service and increase in accident rates. The increasing demand for urban road networks and the spectrum of problems has laid a great setback on traditional traffic management practices. Taking into account the dearth in space and budget constraints and environmental concerns building new and bigger roads is not the answer to solve the present transportation problems. Instead, transportation experts are now focusing on promoting more efficient use of existing capacity of the urban transport network.

Earlier methods involved the usage of probe vehicles for the collection of travel speed data. This was the most common technique used and involved collecting travel time data. Although simple this technique involves certain drawbacks like accuracy varies from technician to technician and there are possibilities of missing checkpoints or inaccurately marked checkpoints. Recent research has demonstrated the feasibility of using Global Positioning System (GPS) and Geographic Information System (GIS) technologies for automating the travel time data collection, reduction, and reporting when using a probe vehicle.

Determining LOS for urban street is very much important as it is the first step of LOS analysis procedure. This affects the planning, design and operational aspects of transportation projects. India being a developing country the traffic is heterogeneous with various kinds of vehicles having various operational characteristics. Presently there is a lack of proper methodology for the determination of LOS for heterogeneous traffic condition. Hence it is barely necessary to determine the LOS in Indian context.

1.3 Objectives and Scope

Based on the above problem statement, the objectives of the study are:

- To classify the urban street segments into various classes using free flow speed data acquired through GPS and data clustering technique.
- To define free flow speed ranges of urban street classes and speed ranges of LOS categories using various clustering algorithms.

1.4 Organization of the Report

This report comprises of seven chapters. The first chapter provides an introduction to this research and also describes the objective and scope of this study. Second chapter deals with discussion on various literatures related to the level of service and use of GIS-GPS in traffic data collection. The third chapter gives idea about the study area of this work and methodology of data collection. The fourth chapter says about the various clustering techniques and cluster validation measure used for this research work. Fifth chapter provides information about the result and analysis. The sixth chapter gives the summary of this study and conclusion of this work. Limitations in the current study and scope for future work are illustrated. References and Appendix are provided at the end of the report.

Chapter 2

Review of Literature

2.1 General

The Level of Service (LOS) concept was first introduced in the 1965 Highway Capacity Manual (TRB, 1965). In the 1965 Highway Capacity Manual (HCM) LOS was stated as "qualitative measure of the effect of numerous factors, which include speed and travel time, traffic interruptions, freedom to maneuver, safety, driving comfort and convenience, and operating cost." After the introduction of the concept of level of service extensive studies were conducted to evaluate the quality of road service as perceived by the users. In 1985 HCM two significant factors i.e. "Qualitative major of operational factors" and "Perception of motorist and passengers" were introduced however "Operational Cost was dropped". In this Highway Capacity Manual version 1965, the level of service was described by the six classes from "A" to "F" defined, based on the combination of travel time and the ratio of traffic flow rate to the capacity, because travel time was recognized as a dominant factor of the service quality. However, these classes were not defined in a quantitative manner. Hence, this concept was redefined in relation to several traffic conditions in the HCM of version 1985 (TRB, 1985). The measures of levels of service adopted in TRB (1985), which describe the characteristics of traffic conditions under operation, include travel speed, traffic flow rate, and traffic density, for each type of road. With the change in the number of vehicles on the road, the amount of congestion, vehicle performance characteristics and geometric standards the concept of Level of Service (LOS) is also changing. What we are following now is the LOS defined in HCM 2000(TRB, 2000). Highway Capacity Manual (HCM, 2000) defined LOS as "a quality measure describing operational conditions within a traffic stream, generally in terms of such service measures as speed and travel time, freedom to maneuver, traffic interruptions, and comfort and

convenience.". The HCM (2000) also designates six levels of service from "A" to "F," for each type of facility, with LOS "A" representing the best operating conditions and LOS "F" the worst. To derive LOS criteria various studies have been performed. Cameron (1996) and Baumgaertner (1996) extended criteria from "A" to "J" and "A" to "I", respectively. Their common concern however was that longer delay occurs due to increasing congestion. However, criteria representing traffic conditions beyond LOS "F" proposed by adding extra categories appears to be somewhat arbitrary. Heterogeneity of traffic, speed regulations, frequency of intersections, presence of bus stops, on-street parking, roadside commercial activities, pedestrian volumes etc are taken as factor affecting the LOS of urban streets (IRC, 1990).

Rigorous research has been carried out to find out various alternatives and solutions for defining and determining the LOS. Kita and Fujiwara (1995) confirmed LOS not just as a traffic operating condition but tried to find the relationship of LOS with driver's perception. Different type of road section has different type of measures of effectiveness, hence it is impossible to evaluate and compare the LOS of different road segments with varying characteristics. Baumgaertner (1996), Cameron (1996) and Brilon (2000) all provided some insight into the limitations of the current LOS measure. Baumgaertner pointed out that continuous growth of urban populations, vehicle ownership, average trip length, and number of trips has resulted in a significant increase in traffic volumes. Commuters have become more adapted to urban congestion, hence the traffic condition that appeared intolerable in 1960, is now considered normal. Spring (1999) objected LOS being a step function. He established service quality being a continuous and subjective matter so it is not wise to use a distinct boundary or threshold value for determining a particular level of service. Shao and Sun (2010) introduced a new concept on LOS. The author divided LOS into two parts: Level of facility supply and Level of traffic operation. Travel speed to free flow speed ratio was considered as evaluation index of traffic operation. Kittelson and Roess (2001) mentioned that the current methodology of determining LOS is not based upon user perception. Clark (2008) objected about the presence of LOS "F" stating it to be very broad. He suggested for a new LOS to be termed as F+ or G. His study specially refers to the type of traffic condition prevailing in New Zealand. Fuzzy set was used by authors to categorize traffic operation into different groups. In July 2001, at the mid-year meeting of the HCQS committee, a motion was passed that stated "The Committee recognizes that there are significant issues with

the current LOS structure and encourages investigations to address these issues" (Pecheux et al., 2004). Flannery et al. (2005) while relating quantitative to qualitative service measuring methods for urban streets found that LOS calculated by HCM (2000) methodology, predicted 35% of the variance in mean driver rating. The authors have suggested that LOS does not completely represent drivers' assessments of performance of urban streets because drivers perceive the quality of urban street segments in several dimensions, including travel efficiency, sense of safety, and aesthetics. Brilon, and Estel, (2010) have presented standardized methods that allow a differentiated evaluation of saturated flow (LOS F) conditions beyond a static consideration of traffic conditions in German Highway Capacity Manual. All the methods described can theoretically be adopted for the evaluation of operational quality on urban roads as well as for signalized and priority intersections.

Arasan and Vedagiri (2010) used computer simulation to study the effect of dedicated bus lane on the LOS of heterogeneous traffic condition in India. The probable modal shift by the commuter was also estimated when a dedicated bus lane was introduced. Flannery et.al. (2008) used user perception to estimate LOS of urban street facilities, for which they used a set of explanatory variables which describe the geometry and operational effectiveness. Chakroborthy and Kikuchi (2007) utilized Fuzzy set to find the uncertainty associated with the LOS categories. Six frameworks were proposed by the authors to determine the uncertainty associated under each LOS category. Cavara et.al. (2011) did not find the traditional algorithms to be much suitable for the analysis of large amount of speed data. Hence, they developed a state of the art hybrid algorithm for this purpose and classified urban roads based on vehicle track and infrastructural data collected through GPS. Chung (2003) tried to find out the travel pattern along a particular route of Tokyo metropolitan area. Small to large ration clustering algorithms were used to cluster historical travel time data for this purpose. The research revealed that day time travel pattern can be classified into three categories i.e. weekdays, Saturday and Sunday (including holidays) but night time travel pattern did not had any such group to classify.

Dandan et.al, (2007) did not take into consideration traffic flow as the only parameter to access the LOS of various traffic facilities. The researcher analyzed the pedestrian LOS with user perception along with physical facilities and traffic flow operation. According to the authors the

primary factors for the classification of LOS can be determined by utilizing mass survey data and statistical software. Fang and Pechuex (2009) studied the LOS of a signalized intersection considering user perception. The researcher used unsupervised data clustering technique such as fuzzy c-means to get distinct cluster of user perceived delay and service rating. The author found it suitable to differentiate LOS into six categories as described in HCM but proposed new six levels of service by emerging LOS A and B and splitting existing LOS F into two categories. Ndoh and Ashford (1994) used fuzzy set theory to develop a model evaluate airport passenger service quality. The authors tried to include user perception in evaluation of service quality instead of just considering traffic parameters for this purpose. Pattnaik and Ramesh Kumar (1996) taking into consideration users' perception tried to develop methodology to define level of service of urban roads.

2.2 GIS and GPS for the Collection of Highway Inventory Data

The most widespread practice to collect speed data is the floating car method. In this method as a driver drives the vehicle a passenger records the elapsed time information at predefined check points. This recording of elapsed time can be done by pen and paper, audio recorder or with a small data recording device. The advantage of this method is it requires very low skilled technician and equipments. The negative aspect of this method is that it is labor intensive and is susceptible to human error. With the development of the computers Distance Measuring Instrument replaced floating car method. DMI measures the speed distance using pulses from a sensor attached to the test vehicle's transmission (Quiroga and Bullock, 1998). This method also has some limitations like the very complicated wiring is required to install a DMI unit to a vehicle. Frequent calibration and verification factors unrelated to the unit are necessary to store making the data file large and which leads to a data storage problem. Current study establishes the usage of GPS receiver in recording location as longitude-latitude, travel time and travel speed. However, additional tools are required to provide a linear reference to these point locations. Fortunately, Geographical Information System (GIS) can be used for this purpose, with the added advantage that the resulting parameters can be entered directly into the existing geographic database (Kennedy, 2002).

Taylor et al. (2000) have developed an integrated Global Positioning System and Geographical Information System for the collection of on-road traffic data from probe vehicle. The system was further integrated with the engine management system of a vehicle to provide time-tagged data on GPS position and speed, distance traveled, acceleration, fuel consumption, engine performance, and air pollutant emissions on a second-by-second basis. Owusu et. al. (2006) tried for management of vehicular traffic in urban road by the data acquired through GPS. They found GIS-GPS environment quite efficient in handling large amount of traffic data and also they tried to give idea to a planner to know the speed which is not acceptable to the driver in congested traffic situation using GPS data.

Cesar and Bullock (1998b) described a new methodology for performing travel time studies using Global Positioning System and Geographic Information System technologies. In this study, the authors have documented the data collection, data reduction, and data reporting procedures using GPS and GIS technologies. For data collection the authors have used GPS receivers to automatically collect coordinates of logged points, travel time, and travel speed at regular sampling periods, for example every one second. Global Positioning Systems (GPS) are one of several available technologies which allow individual vehicle trajectories to be recorded and analyzed. McNally et. al. (2002) designed a flexible GPS-based data collection which incorporates GPS, data logging capabilities, two-way wireless communications, and a user interface in an embedded system which eliminates driver interaction. They also found it suitable to have idea about the individual vehicle trajectory data which is required to get the idea of travel demand at network level. Atikom et. al. (2011) proposed a technique to identify road traffic congestion levels from velocity of mobile sensors with high accuracy and consistent with motorists' judgments. Human perceptions were used to rate the traffic congestion levels into three levels: light, heavy, and jam.

After obtaining the speed data using the GPS-GIS tools the next aim is to classify these data into different groups. For this purpose various clustering techniques can be used.

2.3 Method of Cluster Analysis

2.3.1 Introduction

Clustering methods are applied on the speed data for classification. Various clustering algorithms namely Clustering Large Applications (CLARA), self Organizing Tree Algorithm (SOTA), Hard Competitive Learning (hardcl) and Neural Gas (ngas) have been used for this study purpose. Brief description on some previous research regarding these clustering methods are presented below.

2.3.2 Clustering Large Applications (CLARA)

Boomija (2008) found CLARA to be effective in dealing with large data sets. Murugavel et al. (2011) combined three partition algorithms (PAM, CLARA and CLARANS) with distance based method for outlier detection. Azimi and Zhang (2010) have applied three pattern recognition methods (*k*-means, fuzzy C-means and CLARA) to classify freeway traffic flow conditions on the basis of flow characteristics.

2.3.3 Self Organizing Tree Algorithm (SOTA) Clustering

Wang et.al. (1998) used SOTA to construct phylogenetic trees from biological sequences based on the principles of Kohonen's Self Organizing maps and Fritzke's growing cell structures. Herrero et.al. (2001) used SOTA for the analysis of gene expression data coming from DNA experiments, using an unsupervised neural network. Mateos et.al. (2002) compared various applications of supervised and unsupervised neural networks to the analysis of the gene expression profiles produced using DNA microarrays. The relative efficiencies of different clustering methods for reducing the dimensionality of the gene expression profile data set were studied and SOTA was found to be a good choice for this task.

2.3.4 Hard Competitive Learning (hardcl) Clustering

Dimitriadou et.al. (2004) compared the efficiency and power of several cluster analysis techniques an fully artificial and synthesized fMRI data sets. Yang et.al. (2010) used hard, soft and fuzzy learning schemes for segmenting the ophthalmological MRI data for reducing medical image noise effect with a learning mechanism.

2.3.5. Neural Gas (ngas) Clustering

Martinetz et.al. (1993) used Neural Gas for data compression technique. Neural Gas exhibits good performance in reaching the optimum. Camastra and Vinciarelli (2003) used Neural Gas (NG) along with Learning Vector Quantization (LVQ) for cursive character recognizer. NG is used to verify whether lower and upper case version of a certain letter can be joined in a single class or not. Vascak (2009) used Neural Gas (NG) networks whose role is creating of topologies for complex objects like road networks of municipal communications where it is necessary to determine relations among individual elements.

2.4 Summary

Literature review regarding LOS, speed data collection methods and four clustering algorithms were done. Literature review reveals that there exists certain loop holes in the present LOS methodology given in HCM 2000. Cluster analysis was found to be suitable method for classification of FFS data and average travel speed data for defining speed ranges for urban street classes and speed ranges of LOS categories respectively. GPS has proved to be a very powerful tool in collection of large amount of speed data with utmost accuracy.

Chapter 3 discusses about the Study area and data collection technique for this study.

Chapter 3

Study Area and Data Collection

3.1 Introduction

In the previous chapter details of the literatures are cited. In this chapter, details of study area, data collection and the database preparation are described. The study area for this study is taken as Greater Mumbai. The required GIS layers of the study area for present study obtained from a secondary source. The following section describes the details of road networks and route systems on which the probe vehicle fitted with a Geo-XT GPS receiver was made to run several times. Type and timing of data collection, data smoothing and data compilation are also discussed in detail. Secondary source of data has been used for this study purpose.

3.2 Study Corridors and Data Collection

3.2.1 Base map preparation

The GIS layers of the study area for present analysis are obtained from a secondary source. A detailed roadway inventory survey was also carried out for preparing the digitized GIS base map of the road network.

3.2.2 Study corridors

Five important road corridors of the city of Mumbai in Maharastra state, India are taken up for the present study. Grater Mumbai (GM) is an Island city with a linear pattern of transport network having predominant North-South commuter movements. Passengers move towards south for work trip in the morning hours and return back towards the north in the evening hours. Hence, four north-south corridors and one east-west corridor have been chosen for this study. Major roads like Eastern express highway extending up to south (Corridor-1), LBS Road

extending up to south via Ambedkar road (Corridor-2), Western express highway extending up to marine drive (Corridor-3), SV road extending up to south via Veer Savarkar road (Corridor-4) and Versova- Andheri- Ghatkopar- Vashi (VAGV) (Corridor-5) are included. These five corridors overlapped on the GIS base map of Greater Mumbai are shown in Figure 4.1. These five road corridors as a whole cover 100 street segments with 101 signalized intersections. These corridors were selected from such a large road network, as these have varying road geometric characteristics. Total length of road included in this study is approximately 140 kilometers. Roadway width varies significantly from location to location and in this study it includes two lane undivided roads to eight lane divided roads. Traffic movements are two-way on almost all segments except very few on which traffic movement is restricted to a single direction only. Direction-wise roadway capacity expressed in terms of passenger car unit per hour (PCU/hr) on these road corridors varies from 1500 to 3000. Similarly, free flow speeds (FFS) have a large variation on the road corridors i.e. on some segments FFS is 90 km/hr (Kilometer per hour) and on some segments FFS is mere 25 km/hr. The travel speed also varies widely, which ranges from 5km/hr on some road corridors with highly congested traffic during peak hours to 75 km/hr on corridors during off-peak hours.

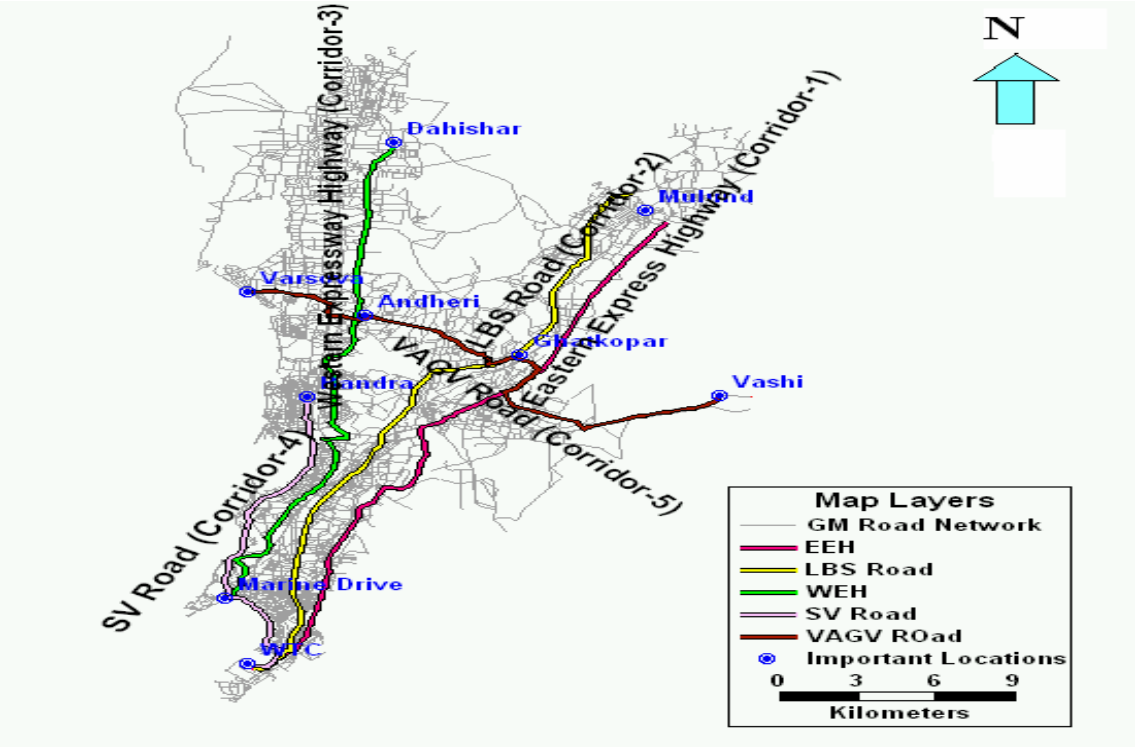


Figure 3.1 Map showing selected corridors of greater Mumbai

3.2.3 Data collection

From extensive literature survey it was found that midsized vehicle fitted with GPS is most suitable for data collection in this study. Midsized vehicle reflect the average of the wide range in vehicle sizes for the heterogeneous traffic flow condition on urban streets in Indian context. It was convenient to employ mid-sized vehicle for this kind of study in which large data samples were collected. Time and resources were the major constraints in the use of different vehicle types in this study. Hence, probe vehicles used in this research work were mid-sized vehicles. These vehicles were fitted with Trimble Geo-XT GPS receiver, capable of logging speed data continuously (at time intervals of one second). The GPS data provides both spatial and time/distance based data from which various traffic parameters were derived, including travel time, stopped time, travel speeds (instantaneous and average). In order to get unbiased data sets, three mid-sized vehicles and the help of three drivers was taken on different days of the survey in this study.

Basically three types of data sets were collected such as roadway inventory details, free flow speed and travel speed during peak and off peak hours. A GPS receiver logs information in the form of features and attributes. A feature is a physical object or an event in the real world for which we want to collect position and descriptive information. A feature is of point, line or area type. We can define a set of attributes for each feature type. An attribute is a piece of descriptive information about the feature. Attributes are of menu, numeric, text, date and time types in data dictionary. Segment number, number of lanes on roadways, median type, parking conditions, pedestrian activity, road side development, access density, commercial activity and speed limits etc. were collected during inventory survey.

The second type of survey conducted was to find the free flow speeds on all these corridors. Before going for the free flow speed data collection, we need to know the time period during which traffic volume would be less than or equal to 200 vehicles per lane per hour. Detailed 24 hour traffic volume count survey was conducted in the month of April; 2005.The traffic volume data were collected on 45 counting stations on seven screen lines covering the whole of Greater Mumbai region. From this survey data, traffic volume per lane per hour was calculated for those roads covered on these five corridors. From this data, it was found that free flow traffic condition (less than 200veh/ln/hr) is approaching at 12 mid-night and all road sections were having free flow traffic conditions from 1 AM to 5 AM. Hence, free flow speed on all these corridors were collected using GPS receiver fitted in probe vehicles during these hours. The third type of data collected was congested travel speed. Congested travel speed survey was conducted during both peak and off-peak hours on both directions of travel on all the corridors. 10-12 travel runs were made on all these five corridors consisting of 100 street segments.

3.3 Summary

Details of the study area, data collection and database preparation have been illustrated in this chapter. The details of corridors on which GPS data was collected were discussed. It was also discussed how the timing of free-flow speed data collection was fixed based upon the traffic volume data. The next chapter gives idea about the cluster analysis algorithms used in this study and also about the various cluster validation parameters used in the research work in order to determine the optimal number of cluster and to select the best clustering algorithm.

Chapter 4

Cluster Analysis

4.1 CLARA Algorithm

4.1.1 Introduction

CLARA (Clustering LARge Applications) was first described by Kauffman and Rousseeuw in 1986. It is constructed especially to cluster large data sets. The clustering of a set of objects with CLARA is carried out in two steps. First a sample is drawn from the set of objects and clustered in to *k* subsets using the *k*-medoid method which also gives *k* representative objects. Then each object not belonging to the sample is assigned to the nearest of the *k* representative objects. This yields a clustering of the entire data sets. A measure of the quality of this clustering is obtained by computing the average distance between each object of the data set and its representative object. After random samples have been drawn and clustered, the one is selected for which the lowest average distance was obtained. The resulting clustering of the entire dataset is then analyzed further. For each cluster CLARA gives its size and its medoid and prints a complete list of its objects. Also a graphical representation of the clustering is provided by means of the silhouettes.

4.1.2 Details of Clustering Large Applications (CLARA) algorithm

The clustering of a set of data points is carried out in two steps. First, a sample is drawn from the data and is clustered into *k* clusters using the *k*-medoid method [PAM], which gives *k* representative points (medoids) of the data set. Then each data point which does not belong to the whole data set. A quality measure of this partition is obtained by calculating the average distance between each data point of the data set and its center (medoid). After all the samples have been drawn and clustered, the one with the lowest average distance is obtained. The

clustering for the sample takes in two phases. At first, an initial clustering is obtained by the successive selection of representative points until a number of *k* ones has been found.

1. A point x_i is chosen. (not yet selected)
2. A non-selected x_i point and calculated the difference between its dissimilarity D_j with the most similar previously selected one, and its dissimilarity $d(i, j)$ with the point x_i.
3. If the difference is positive, the x_j contributes to the decision to select

$$x_i.C_{ij} = \max\left(D_j - d(j,i),0\right) \quad (1)$$

4. The total gain by selecting x_i is: $\sum_j C_{ji}$ (2)
5. x_i is chosen which maximizes $\max_i \sum_j C_{ji}$ (3)

In the second phase, the set of representative points is improved by considering all pairs of points (x_i, x_i) for which point x_i has been selected and point x_i has not. Thus, this is the effect on the value of clustering when a swap takes place.

Figure 4.1 shows the steps in which CLARA algorithm is executed.

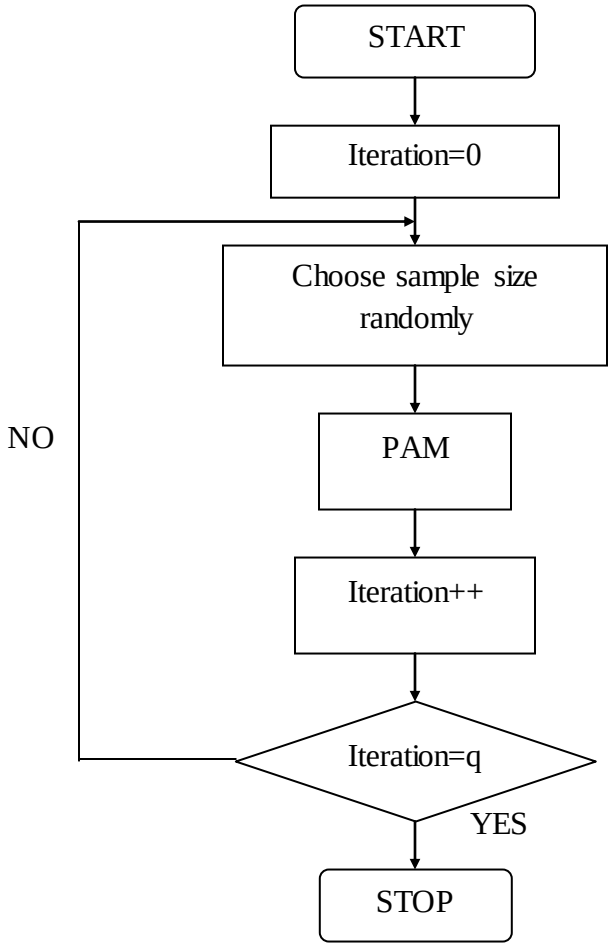


Figure 4.1 Flowchart of CLARA Clustering

4.2 Self Organizing Tree Algorithm (SOTA)

4.2.1 Introduction

Self-Organizing Tree Algorithm (SOTA) was developed by Dopazo and Carazo in 1997. SOTA is an unsupervised neural network with a binary tree topology. SOTA is a divisible method i.e. the clustering process is performed from top to bottom, i.e. highest hierarchical levels are resolved before going to the details of the lowest levels. It combines the advantages of both hierarchical clustering and Self-Organizing Maps (SOM). The algorithm picks a node with the latest diversity and splits it into two nodes, called cells. This process can be stopped at any level,

assuming a fixed number of hard clusters. This behavior is achieved with setting the unrest growth parameter to TRUE. Growth of the tree can be stopped based on other criteria, like the allowed maximum diversity within the cluster and so on. Since SOTA runtimes are approximately linear with the number of items to be classified, it is especially suitable for dealing with huge amounts of data. The method proposed is very general and applies to any data providing that they can be coded as a series of numbers and that a computable measure of similarity between data items can be used.

4.2.2 Details of Self Organizing Tree Algorithm (SOTA)

The pseudocode of SOTA algorithm is as follows:

Step 1:

The system is initialized.

Step 2:

New input is presented.

Step 3:

Distance to all external units is computed. For aligned sequences, distance between input sequence j and the unit i is computed as:

$$d_{s_i c_j} = \sum_{j=1}^{L} \frac{1 - \sum_{i=1}^{A} S_j[r,l] C_i[r,l]}{L} \tag{4}$$

Where $S_j[r,l]$ is the value for the residue r of the input sequence node j and $C_i[r,l]$ is the residue r of the neuron i. Output unit is selected i^* with minimum distance d_{ij}.

Step 4:

Unit i^* is updated and neighbours neurons updated as:

$$C_i(T+1) = C_i(T) + \eta_{T,i,j}\left(S_j - C_i(T)\right) \tag{5}$$

Where $\eta_{T,i,j}$ is the neighborhood function for unit I.

Step 5:

If a cycle is finished, the size of the network is increased i.e. two new units are attached to the original unit with higher resources. This unit becomes the mother unit and does not receive any more updating.

Resources for each terminal unit *i* are calculated as an average of the distances of the input sequences assigned to this unit itself.

$$R_i = \frac{\sum_{k=1}^{K} d_{S_k C_i}}{k} \tag{6}$$

Step 6:

The above process is repeated by going to Step 2 until convergence.

Figure 4.2 shows the steps in which SOTA algorithm is executed.

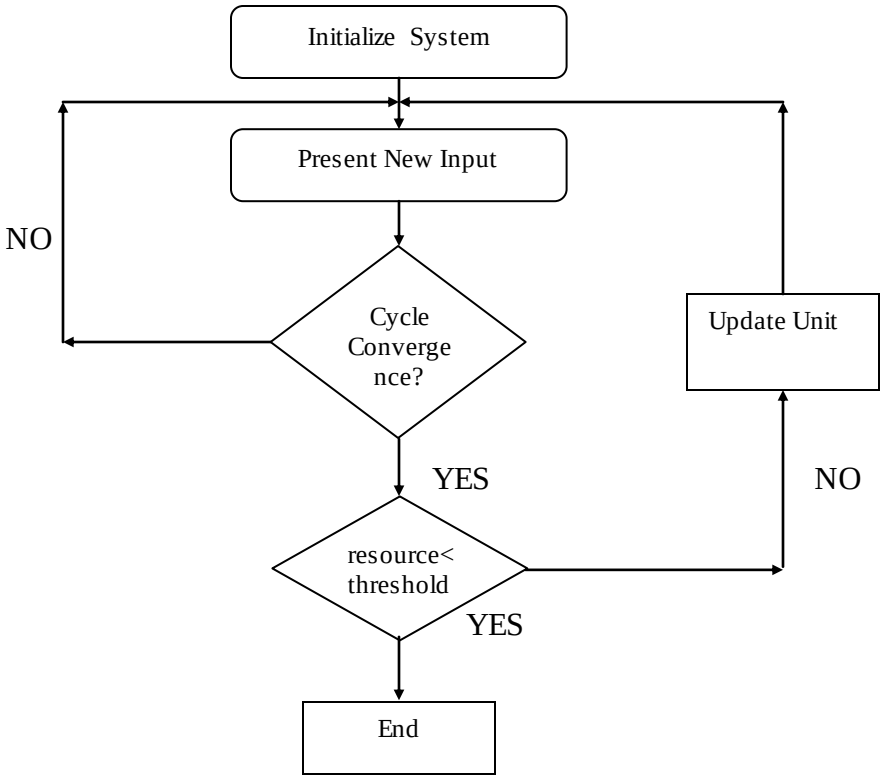


Figure 4.2 Flowchart of SOTA Clustering

4.3 Hard Competitive Learning (*hardcl*) Algorithm

4.3.1 Introduction

Hard competitive learning is a well known online stochastic gradient descent algorithm for minimization of the average distance of a given set of data to its closest center. It was proposed by Fritzke, B. (1997). *hardcl* is the simplest online clustering algorithm, where only one output unit (the cluster center) is the winner (the closest to the data point) for each given data point, and the vector of the winner moves towards the vector of the given point. For *hardcl* the set of cluster centers $C = \{c_1, c_2, ..., c_n\}$ are seen as output units of a neural network, which adapt every time a new data point is presented.

4.3.2 Details of Hard Competitive Learning (*hardcl*) algorithm

The pseudocode of *hardcl* algorithm is presented below:

Step 1:

The set C to contain $k(k << N)$ units $c_j : C = \{c_1, c_2, ..., c_k\}$, with center vectors $W_{cj} \in \Re^d$ chosen randomly from the data set is to be initialized. The iteration count is set to zero.

Step 2:

A pattern x_i is drawn from the data set.

Step 3:

The winner $S(x_i) : S(x_i) = \arg\min \left\|_{c \in C} x_i - W_c \right\|$ is determined. (7)

Step 4:

The center vector of the winner is moved along the gradient of $\left\| x_i - W_{S(x_i)} \right\|$ toward x_i. In the case of the Euclidean norm this is $\Delta W_{S(x_i)} = \varepsilon_t \left(x_i - W_{S(x_i)} \right)$, where ε_t is suitable chosen learning rate.

Step 5:

t is set as $t := t + 1$;if $t < t_{\max}$,else return to step 2.

4.4 Neural Gas Algorithm

4.4.1 Introduction

The algorithm was proposed by Martinetz and Schulten (1991). Neural gas algorithm is a clustering algorithm which allows to find a generalization for a set of data represented as a group with similar characteristics. The task of clustering can be realized in many ways for example finding the number of groups, clustering a set of n-groups. Each group is a vector of features.

4.4.2 Details of Neural Gas Algorithm

The pseudo code of the neural gas algorithm is as follows:

Step 1:

Initialize the set C to contain N units c_j

$$C = \{C_1, C_2, ..., C_k\}$$

with center vectors $W_{cj} \in R^d$ chosen randomly from the data set. Set the iteration counter t to zero.

Step 2:

Draw an item x_i from the data set.

Step 3:

Order all elements of C according to their distance x_i, i.e., find the sequence of indices $(l_0, l_1, ..., l_{k-1})$ so that the center vector W_{l0} is closest to the pattern, W_{l1} the second closest and etc. We denote the number associated with W_l as $m_l(x_i C)$.

Step 4:

Adapt the center vectors according to

$\Delta W_1 = \varepsilon_t h_\lambda (m_l(x_i, C)) v \| x_i - w_l \|$ where,

$$\lambda_t = \lambda_{ini} \left(\lambda_{fin} / \lambda_{ini} \right)^{\left(t / t_{\max} \right)} \quad (8)$$

$$\varepsilon_t = \varepsilon_{ini}\left(\varepsilon_{fin}/\varepsilon_{ini}\right)^{(t/t_{\max})} \tag{9}$$

$$h_\lambda(m) = \exp\left(-m/\lambda_t\right) \tag{10}$$

and $\nabla\|(x_i - w_i)\|$ denotes the gradient of the distance function.

Step 5:

Set $t := t+1$; if $t < t_{\max}$ continue with step 2

Suitable initial values $(\lambda_{ini}, \varepsilon_{ini})$ and final values $(\lambda_{fin}, \varepsilon_{fin})$ have to be chosen for the iteration dependent parameters λ_t and ε_t.

4.5 Cluster Validation Measures

Quality of the clustering result obtained from a clustering algorithm can be checked by various cluster validation measures. These Validation parameters mainly used to evaluate and compare whole partitions, resulting from different algorithms or resulting from the same algorithms under different parameters. Most common applications of cluster validation is to determine the optimal number of clusters for a particular data set (Bensaid et al., 1996). The problems of deciding the number of clusters better fitting a data set as well as the evaluation of the clustering results has been subject of several research efforts (Dave, 1996; Gath and Geva, 1989; Rezaee et al., 1998; Smyth, 1996; Theodoridis and Koutroubas, 1999; Xie and Beni, 1991). Since clustering is an unsupervised method and there is no a-priori indication for the actual number of clusters presented in a data set, there is a need of some kind of clustering results validation. Different validity measures have been proposed in the literature, none of them is perfect by oneself, and therefore several parameters are used in the study.

A) Calinski-Harabasz Index (CHI)

Calinski-Harabasz Index is an index which can be used to get the optimal number of clusters from a particular set of data (Calinski, R.B., Harabasz, J., 1974).

$$GI(u) = \left[\frac{trace(B_u)}{u-1} \right] \Big/ \left[\frac{trace(W_u)}{n-u} \right] \tag{11}$$

Where $x = \{x_{ij}\}$; $i = 1,....,n$; $j = 1,....,m$ – data matrix, n = number of objects, m= number of variables, u=number of clusters $(u = 2,....,n-1)$

(12)

$$W_u = \sum_r \sum_{i \in C_r} \left(X_{ri} - \overline{X_r} \right) \left(X_{ri} - \overline{X_r} \right)^T \tag{13}$$

—within group dispersion matrix for data clustered into u clusters,

$$B_u = \sum_r n_r \left(\overline{X_r} - \overline{X} \right) \left(\overline{X_r} - \overline{X} \right)^T \tag{14}$$

— between group dispersion matrix for data clustered into u clusters,

Where $r = 1,....,u$ — cluster number, $\overline{X_r}$ — centroid or medoid of cluster, $\overline{X}$ — centroid or medoid of data matrix, C_r — the indices of objects in cluster r, n_r — number of objects in cluster r. The value of u, which maximizes GI(u), is regarded as specifying the number of clusters.

B) Connectivity Index (CI) (Handl et al., 2005)

Let N denote the total number of observations (rows) in a dataset and M denotes the total number of columns, which are assumed to be numeric (e.g., a collection of samples, time points, etc.). Define $nn_{i(j)}$ as the jth nearest neighbor of observation i, and let $x_i, nn_{i,(j)}$ be zero if I and j are in the same cluster and 1/j otherwise .

Then, for a particular clustering partition $C = \{C_1, C_2, ..., C_k\}$ of the N observations into K disjoint clusters, the connectivity is defined as

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_i, nn_{i(j)} \tag{15}$$

Where L is a parameter giving the number of nearest neighbors to use. The connectivity has a value between zero and ∞ and should be minimized.

C) Average Proportion of Non-overlap (APN) Index (Datta and Datta, 2003)

The APN measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data in a single column removed. Let $C^{i,0}$ represent the cluster containing observation i using the original clustering (based on all available data), and $C^{i,l}$ represent the cluster containing observation i where the clustering is based on the dataset with column l removed. Then, with the total number of clusters set to K, the APN measure is defined as

$$APN(K) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{(1-n)\left(C^{i,l} \cap C^{i,0}\right)}{n\left(C^{i,0}\right)} \tag{16}$$

The APN is in the interval [0,1], with values close to zero corresponding with highly consistent clustering results.

D) Average Distance (AD) Index (Datta and Datta, 2003)

The AD measure computes the average distance between observations placed in the same cluster by clustering based on the full data and clustering based on the data in a single column removed. It is defined as

$$AD = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,0} \cap C^{i,l})}\left[\sum_{i\in C^{i,0},\, j\in C^{i,l}} dist(i,j)\right] \tag{17}$$

The AD has a value between zero and ∞, and the smaller values are preferred.

E) Average Distance between Means (ADM) Index (Datta and Datta, 2003)

The ADM measure computes the average distance between cluster centers for observations placed in the same cluster by clustering based on the full data and clustering based on the data in a single column removed. It is defined as

$$ADM(K) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}\right) \tag{18}$$

Where $\bar{x}_{C^{i,0}}$ is the mean of the observations in the cluster which contain observation i, when clustering is based on the full data, and $\bar{x}_{C^{i,l}}$ is similarly defined. Currently, ADM only uses the Euclidean distance. It also has a value between zero and ∞, and smaller values are preferred.

F) Figure of merit (FOM) Index (Datta and Datta, 2003)

The FOM measures the average intra-cluster variance of the observations in the deleted column, where the clustering is based on the remaining (un-deleted) samples. These estimates the mean error using predictions based on the cluster averages. For a particular left out column l, the FOM is given by

$$FOM(l,K) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)}\left(x_{i,l},\bar{x}_{Ck,(l)}\right)}, \tag{19}$$

Where $x_{i,l}$ is the value of the i^{th} observation of the l^{th} column in the cluster $C_k(l)$, and $\overline{x}_{c_k(l)}$ is the average of cluster $C_k(l)$. Currently the only distance available in FOM is Euclidean. The FOM is multiplied by an adjustment factor $\sqrt{\frac{N}{N-K}}$, to alleviate the tendency to decrease as the number of clusters increases. The final score is averaged over all the removed columns and has a value between zero and ∞ with smaller values equaling better performance.

G) Homogeneity Index (HI) (Datta and Datta, 2006)

As its name suggests, the HI measures how homogenous the clusters are. Let B={B_1,…..,B_F} be a set of F functional classes, not necessarily disjoint and let B(i) be the functional class containing class i. Similarly, we define B(j) as the function class containing class j and assign the indicator function I(B(i)=B(j)) the value 1 if B(i) and B(j) match and o otherwise. Intuitively, we hope that class placed in the same statistical cluster also belong to the same functional classes. Then, for a given statistical clustering partition C={C_1,…..,C_k} and set of classes B, the HI is defined as

$$HI(C,B) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k(n_k-1)}\sum_{i\neq j\in C_K} I\left(B(i)=B(j)\right) \tag{20}$$

Here $n_k = n(C_K \cap B)$ is the number of annotated classes in statistical cluster C_k. The HI is in the range [0,1], with larger values corresponding to more homogenous clusters.

H) Stability Index (SI) (Datta and Datta, 2006)

The SI is similar to the other stability measures, and inspects the consistency of clustering for genes with similar biological functionality. Each sample removes one at a time, and the cluster

membership for genes with similar functional annotation is compared with the cluster membership using all available samples. The SI is defined as

$$SI(C,B)=\frac{1}{F}\sum_{k=1}^{F}\frac{1}{n(B_k)(n(B_k)-1)M}\sum_{l=1}^{M}\sum_{i\neq j\in B_k}\left(\frac{n\left(C^{i,0}\cap C^{j,l}\right)}{n\left(C^{i,0}\right)}\right) \tag{21}$$

Where F is the total number of functional classes, $C^{i,0}$ is the statistical cluster containing observation I based on all the data, and $C^{j,l}$ is the statistical cluster containing observation j when column l is removed. The SI is in the range [0,1], with larger values corresponding to more stable clusters of the functionally annotated class.

I) Tau Index (Rohlf, 1974)

Tau index is computed as follows:

$$Tau=\frac{S(+)-S(-)}{\left[\left(\frac{n_d(n_d-1)}{2-t}\right)\times\left(\frac{n_d(n_d-1)}{2}\right)\right]^{0.5}} \tag{22}$$

Where

$S(+)$ is the number of concordant comparisons, S(-) is the number of disconcordant comparisons, n_d is the total number of distances (which is the same as the total number of observations or objects under study.), t is the number of comparisons of two pairs of points where both pairs represent within cluster comparisons or both pairs are between cluster comparisons. The maximum value in the hierarchy sequence was taken as indicating the correct number of clusters.

J) Ratkowsky Index (Ratkowsky and Lance, 1978)

The index is based on the formula:

$$\frac{\bar{S}}{q^{1/2}} \tag{23}$$

The value of $\bar{S}$ is equal to the average of the ratios of B/T where B stands for the sum of squares between the clusters for each variable and T for the total sum of squares for each variable. The optimal number of clusters is that value of q for which $\bar{S} / q^{(1/2)}$ has its minimum value.

K) Duda Index (Duda and Hart, 1973)

$$duda = \frac{J_e(2)}{J_e(1)} = \frac{W_k + W_l}{W_m} \tag{24}$$

Where, $J_e(2)$ is the sum of squared errors within the cluster when the data are partitioned into two clusters and $J_e(1)$ gives the squared errors when only one cluster is present.

W_k, W_l, W_m are defined as W_q

$$W_q = \sum_{k=1}^{q} \sum_{i \in c_k} (x_i - c_k)(x_i - c_k)^T \tag{25}$$

is the within group dispersion matrix for data clustered into q clusters.

It is assumed that clusters c_k and c_l are merged to form c_m

$$B_{kl} = W_m - W_k - W_l, \text{ if } c_m = c_k \cup c_l \tag{26}$$

n_i =number of observations in cluster c_i, $i = k.l.m$

The optimal number of clusters is the smallest q such that

$$duda = 1 - \frac{2}{\pi p} - z\sqrt{\frac{2\left(1 - \left(\frac{s}{\pi^2 p}\right)\right)}{n_m p}} = \text{critical value} \tag{27}$$

Where, p=number of variables in the data set

Z=Standard normal score

L) Gplus Index (Rohlf, 1974)

$$G(+) = \frac{2S(-)}{n_d(n_d - 1)} \tag{28}$$

Where,

S(-) is the number of discordant comparisons i.e. the number of times where two points which were in the same cluster had a larger distance than two points not clustered together. n_d =total number of distances (which is the same as the total number of observations or objects under study).Minimum values of the index are used to determine the optimal number of clusters in the data.

M) McClain Index (Milligan and Cooper, 1985)

This index consists of the ratio of two terms. The first term is the average within cluster distance divided by the number of within cluster distances. The denominator value was the average between cluster distance divided by the number of cluster distances. It is computed as follows:

$$mcclain = \frac{mcan\left(\sum_{q}^{k=1}\sum_{i=1}^{n_k}\sum_{j=i+1}^{n_k} d_{ij}\right)}{mcan\left(\sum_{k=1}^{q}\sum_{i\in c_k}\sum_{l=k+1}^{q}\sum_{j\in c_l} d_{ij}\right)} \tag{29}$$

q is the number of clusters, n_k is the number of objects in the k^{th} cluster, $k \in [1....q] k$ in $[1....q]$, d_{ij} = distance between i^{th} and j^{th} objects. The minimum value of the index is used to indicate the optimal number of cluster.

N) PtBiserial Index (Milligan, 1980)

$$Ptbiserial = \frac{\left(\overline{d_b} - \overline{d_w}\right)\left[f_w f_b \big/ \left(n^2 d\right)\right]^{0.5}}{S_d} \tag{30}$$

Where,

d_w = sum of within cluster distances, d_b = sum of between cluster distances, , $\overline{d_b}, \overline{d_w}$ are respective means, S_d = standard deviation of all distances, n_d = total number of distances,

f_w = number of within cluster distances, f_b = number of between cluster distances. Its value varies between -1 to +1 and the maximum values indicate the optimal number of clusters.

4.6 Summary

Details of cluster analysis and various algorithms associated with it are discussed in this chapter. Details of four clustering algorithms i.e. CLARA, SOTA, hardcl and ngas used for this study purpose are discussed. Various cluster validation parameters and their significance in finding the optimal number of clusters for the input data are elaborated.

Chapter 5

Result Analysis for Defining LOS

5.1 Introduction

Results of cluster analysis performed are discussed in this chapter. The four clustering algorithms namely CLARA, SOTA, hardcl and ngas were used on the FFS data to classify the urban street segments into four classes. After the segments were defined in particular classes of urban streets, speed range for six LOS were found out using the same clustering algorithms.

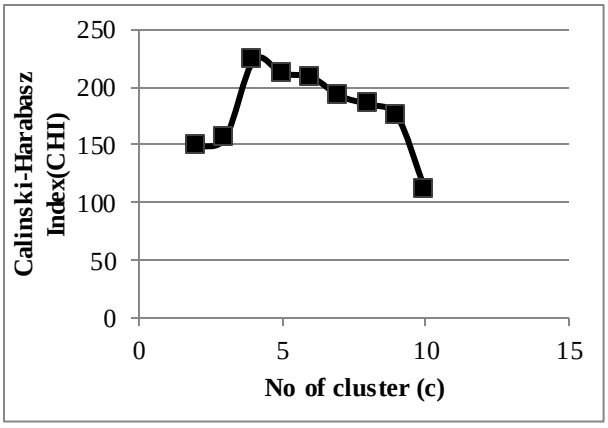
5.2 Application of Cluster Analysis Methods in Defining LOS Criteria of Urban Streets.

Average travel speeds were calculated direction wise on each segment. Four clustering algorithms (CLARA, SOTA, hardcl and ngas) were applied in two stages. First, clustering method was applied on average free flow speeds of all segments and free-flow speeds were classified into four groups. Urban street segments were classified into urban street class of I to IV. Secondly, clustering methods were applied on average travel speeds collected during peak and off peak hours on street segments for each of the urban street classes. In the second case, speeds were classified into six groups (A to F) for six categories of levels of service; thus speed ranges from level of service categories were defined in Indian context.

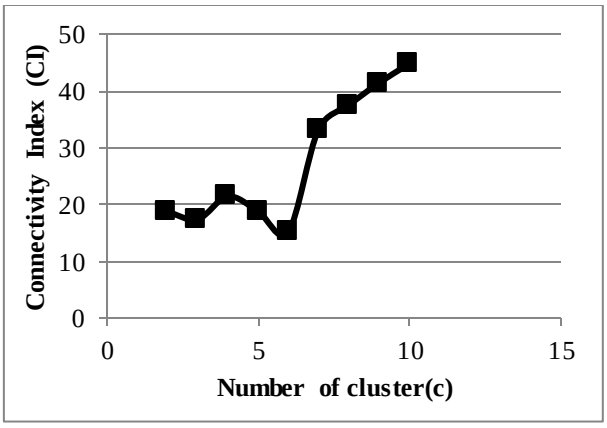
5.2.1 CLARA clustering

The free flow speed data acquired through GPS receiver was clustered using the CLARA algorithm. In this research eight validation parameters were used. Value of validation parameters were obtained for 4 to 6 number of cluster and were plotted in Figure 5.1 (A) to Figure 5.1 (H). These eight numbers of validation parameters were used to know the optimum number of cluster for this particular data set of free flow speed. The optimum number of cluster tells us the number

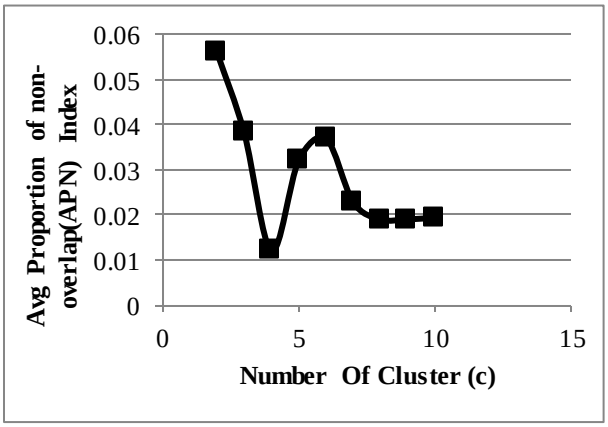
of urban street classes into which the urban street segments should be classified into. According to the available literature the maximum value of Calinski-Harabasz Index (CHI) signifies the optimal number of clusters for a particular data set. Figure 5.1(A) shows that the Index is maximum for 4 numbers of clusters. The available literature says that the minimum value of Connectivity Index (CI) gives the optimal number of clusters. Figure 5.1 (B) shows that the Index are minimum for 6 numbers of clusters. Figure 5.1 (C) shows Average Proportion of Non-Overlap (APN) Index. The value close to zero gives the optimal number of clusters. So, going with the literature four is taken as the optimal number of cluster. For Average Distance (AD) Index, Average Distance between Means (ADM) Index and Figure of Merit Index (FOM) Index the smaller values give the optimal number of clusters. Figure 5.1 (D), (E), (F) shows that the optimal number of clusters is obtained to be 4, 4 and 5 respectively. According to the available literature larger values of Homogeneity Index (HI) and Stability Index (SI) gives the optimal number of clusters. Figure 5.1 (G), (H) shows that the optimal number of clusters to be 4. Out of eight validation parameters considered in this study six parameters give the optimal cluster value as 4 which is also same as suggested by HCM-2000. That is the reason for which in this research the urban street segments were classified into four Classes by using CLARA algorithm.



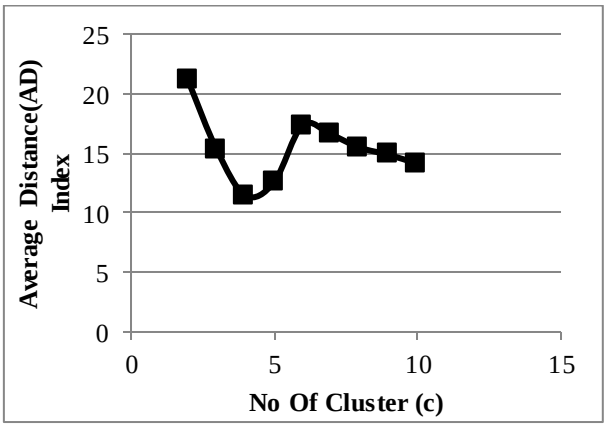
A: CHI vs Number of cluster



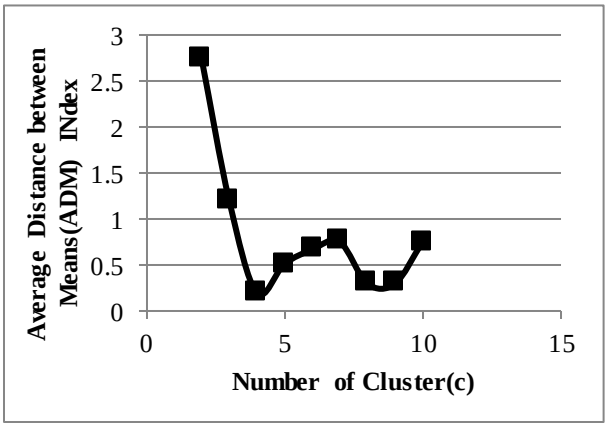
B: CI vs Number of cluster



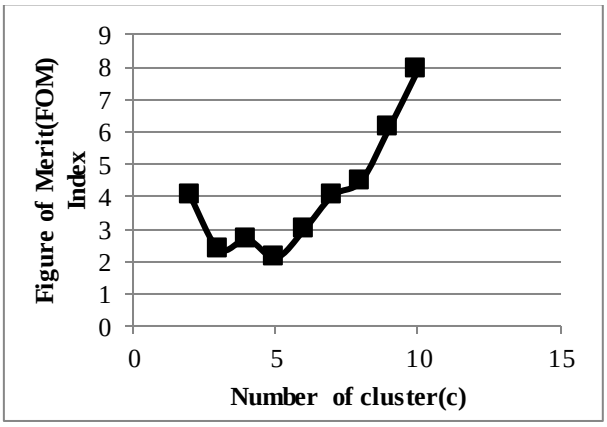
C: APN vs Number of cluster



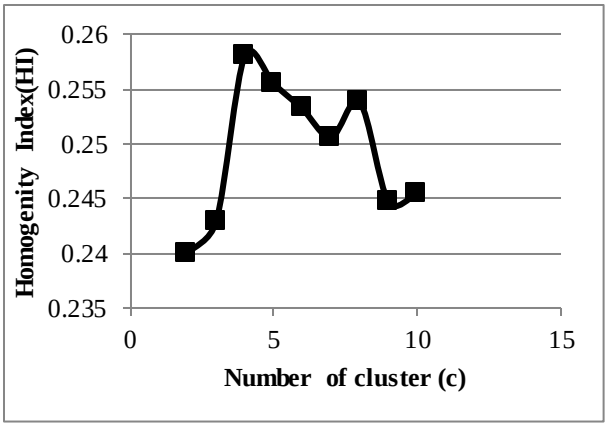
D: AD vs Number of cluster



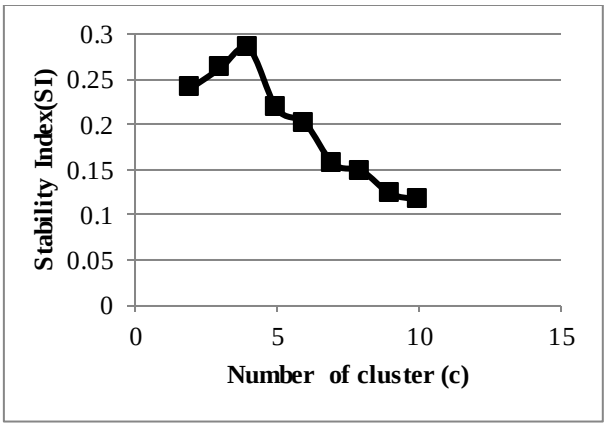
E: ADM vs Number of cluster



F:FOM vs Number of cluster



G: HI vs Number of cluster



H: SI vs Number of cluster

Figure 5.1 Validation measures for optimal number of clusters using CLARA clustering

In this study 100 urban street segments of five urban street corridors were analyzed. So to get the FFS ranges of different urban street class FFS of these 100 urban street segments were clustered using CLARA. Different symbol in the plot used for different urban street class. Figure 5.2 shows the speed ranges for different urban street classes. Observing this figure it can be inferred most of the street segments belong to Urban Street class-I and Urban street class-II. It is observed from the collected data set that when a street segment falls under particular urban street class, it agrees with the geometric and surrounding environmental condition of the road segments as well. It has been found that there is very good correlation exists between free flow speed and geometric and environmental characteristics of streets under considerations.

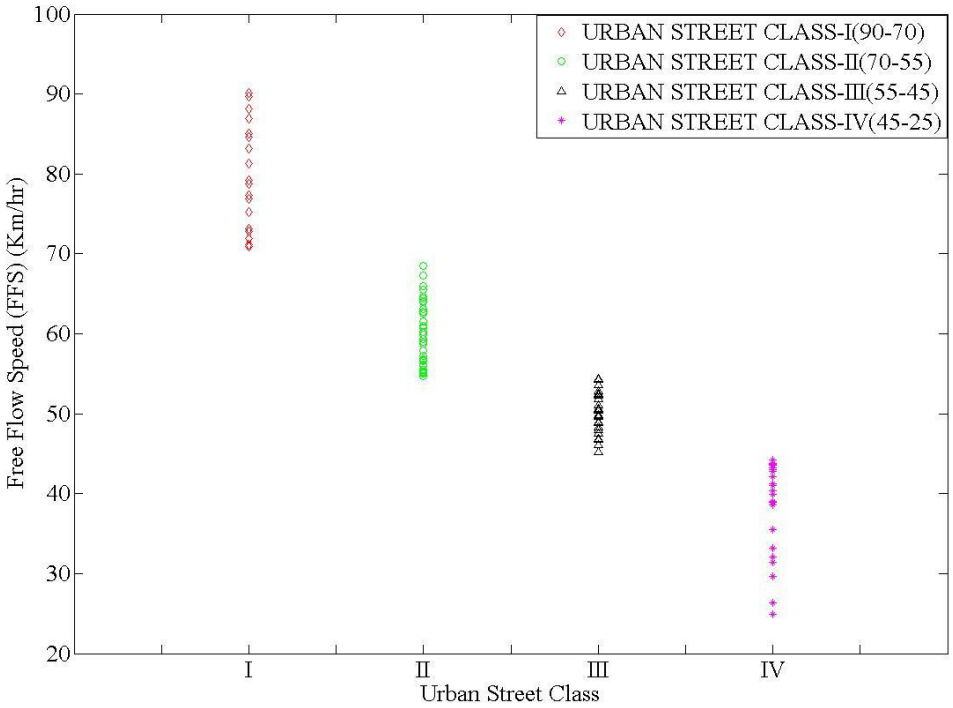
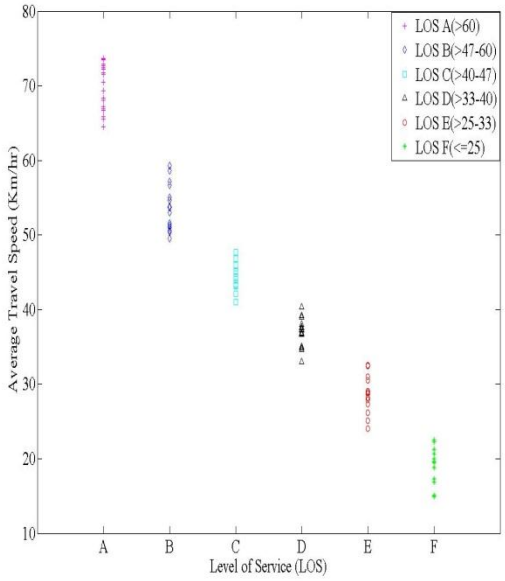
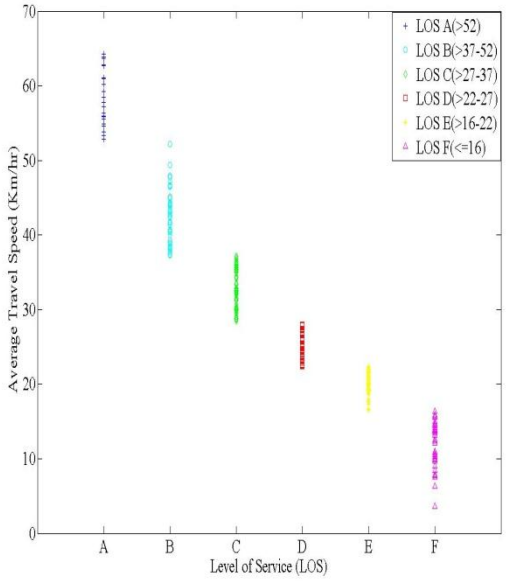


Figure 5.2 CLARA Clustering of FFS for Urban Street Classification

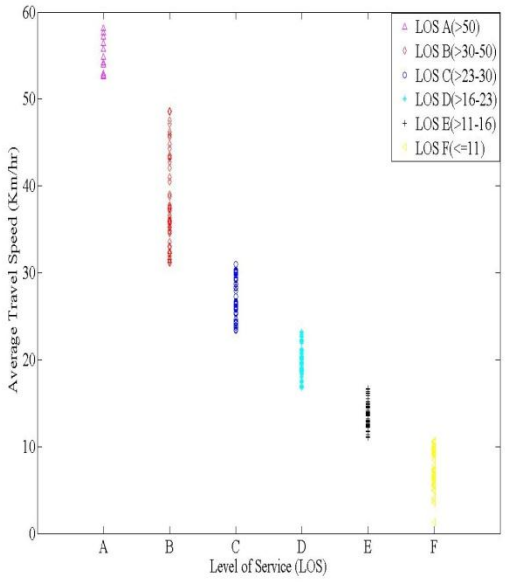
After classification of urban streets into number of classes, direction wise average travel speed on street segments during both peak and off peak hours were clustered using CLARA to find the speed range of level of service categories. In fig. 5.3 the speed values are shown by different symbols depending on to which LOS category they belong. The legends in fig. 5.3 (A-D) gives the speed ranges for the six LOS categories obtained by using CLARA clustering. The speed ranges for LOS categories found using CLARA clustering is also shown in Table 5.1.



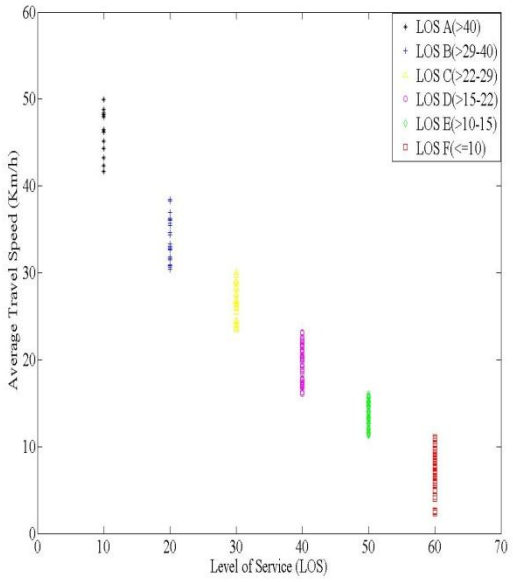
A: LOS of Urban Street Class I



B: LOS of Urban Street class II



C: LOS of Urban Street Class III



D: LOS of Urban Street Class IV

Figure 5.3 Level of service of urban street classes (I-IV) using CLARA clustering on average travel speeds

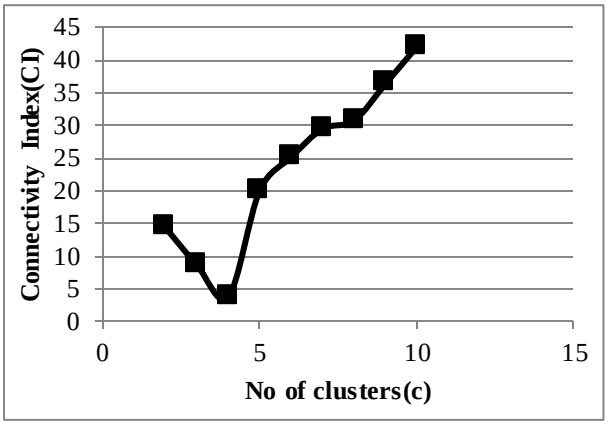
Table 5.1 Urban speed ranges for different LOS proposed in Indian conditions by CLARA method

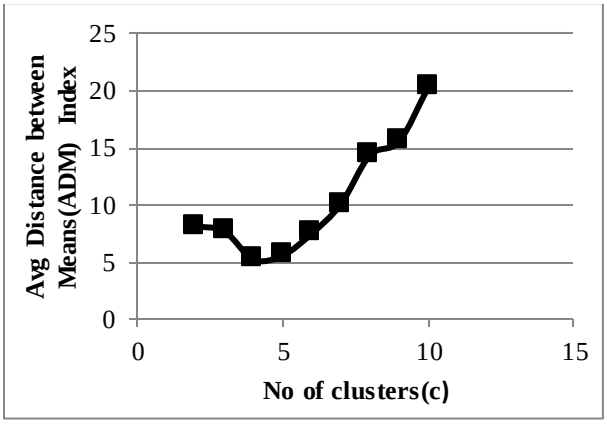
Urban Street Class	I	II	III	IV
Range of Free Flow Speed(FFS)	90 to 70 Km/hr	70 to 55 Km/hr	55 to 45 Km/hr	45 to 25 Km/hr
Typical FFS	80 Km/hr	60 Km/hr	50 Km/hr	35 Km/hr
LOS	Average Travel Speed(Km/hr)			
A	>60	>52	>50	>40
B	>47-60	>37-52	>30-50	>29-40
C	>40-47	>27-37	>23-30	>22-29
D	>33-40	>22-27	>16-23	>15-22
E	>25-33	>16-22	>11-16	>10-15
F	≤25	≤16	≤11	≤10

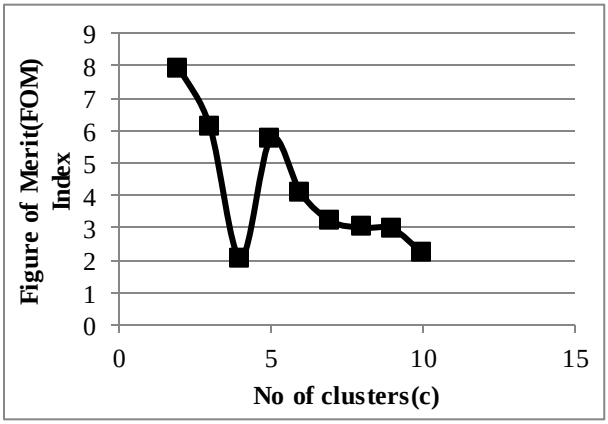
Average travel speed of LOS categories (A-F) expressed in percentage of free flow speeds were found to be approximately 85 and above, 70-85, 55-70, 40-55, 30-40 and less than equal to 30 respectively. Whereas in HCM (2010) it has been mentioned that these values are 85 and above, 67-85, 50-67, 40-50, 30-40 and less than equal to 30 percentage respectively.

5.2.2 SOTA Clustering

SOTA is a clustering tool used in this study purpose to get the speed ranges for different urban street classes and travel speed ranges of LOS categories. Five validation parameters were used for SOTA clustering. The validation parameters used for SOTA clustering are Connectivity Index (CI), Average Distance between Means (ADM) Index, Figure of Merit (FOM) Index, Homogeneity Index (HI), Stability Index (SI), Average Proportion of Non-Overlap (APN) Index and Average Distance (AD) Index. Free flow speed (FFS) was used as input for these validation indices. The optimal numbers of clusters were obtained between 4 to 6. Values of validation parameters obtained are plotted in Figure 5.4 (A) to 5.4 (G).

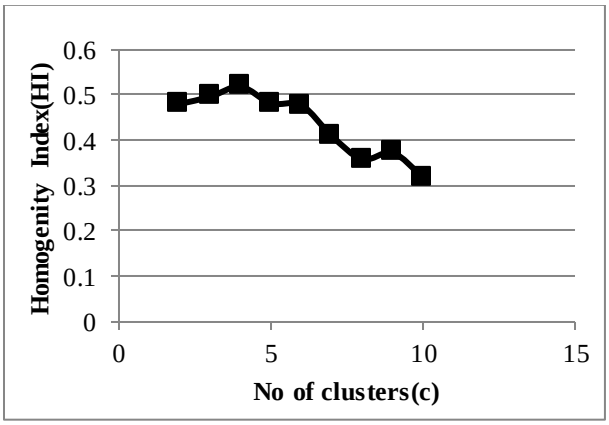






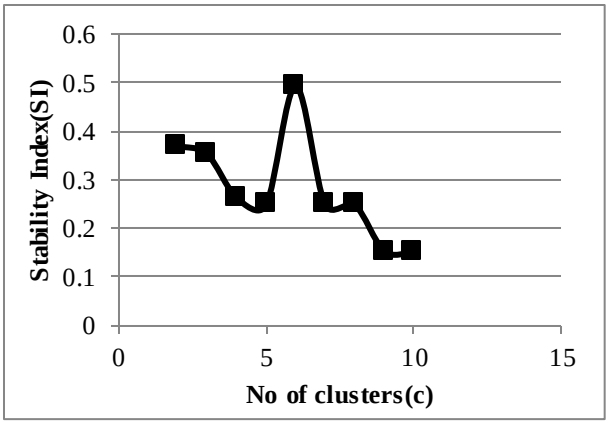
A: CI vs Number of clusters

B: ADM vs Number of clusters

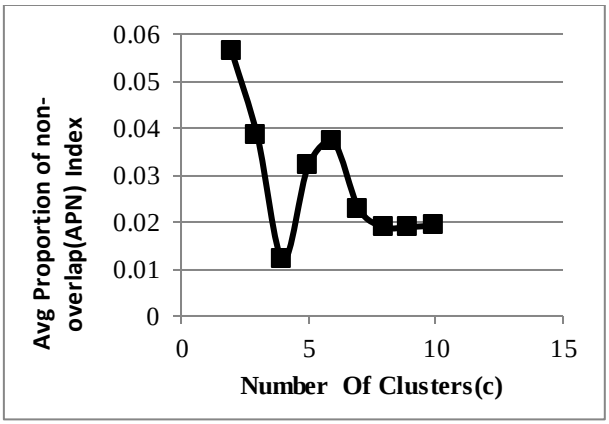


C: FOM vs Number of clusters

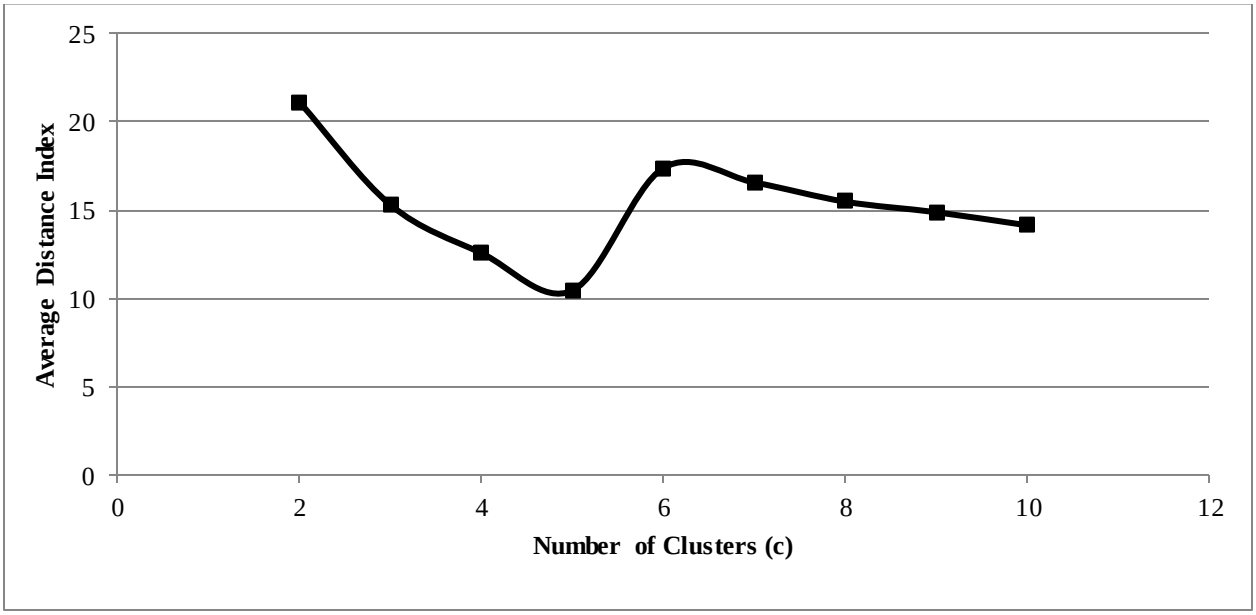
D: HI vs Number of clusters



E: SI vs Number of clusters



F: APN vs Number of clusters



G: AD vs Number of clusters

Figure 5.4 Validation measures for optimal number of clusters using SOTA clustering

These seven validation parameters are used to know the optimum number of clusters for FFS data set. Optimal number of clusters are basically concerned with the quality of the cluster obtained by applying a particular clustering algorithm to a particular data set. Every algorithm

has its natural way of classification of the data set into numbers of groups. When a data set is clustered into its optimal number of clusters the quality of the cluster is best as the variation between the data points belonging to a particular cluster is minimal. The available literature says that the minimum value of Connectivity Index (CI) gives the optimal number of clusters. Figure 5.4 (A) shows that the Index is a minimum of 4 numbers of clusters. For Average Distance between Means (ADM) Index and Figure of Merit Index (FOM) Index the smaller values give the optimal number of clusters. Figure 5.4 (B), (C) shows that the optimal number of clusters are obtained to be 4 respectively. According to the available literature larger values of the Homogeneity Index (HI) and Stability Index (SI) gives the optimal number of clusters. Figure 5.4 (D), (E) shows that the optimal number of clusters to be 4 and 6 respectively. Figure 5.4 (F) shows Average Proportion of Non-Overlap (APN) Index. The value close to zero gives the optimal number of clusters. So, going with the literature four is taken as the optimal number of clusters. Moreover literature review depicts that for Average Distance (AD) Index smaller value gives the optimal number of clusters. Figure 5.4 (G) shows the optimal number of clusters to be 5. Out of seven validation parameters considered in this study five parameters give the optimal cluster value as 4 which is also same as suggested by HCM-2000. That is the reason for which in this research the urban street segments were classified into four Classes by using SOTA algorithm.

Hundred urban street segments of five urban street corridors were analyzed in this research. So to get the FFS ranges of different urban street classes FFS of these 100 urban street segments were clustered using SOTA. SOTA uses both the properties of hierarchical and SOM. Speed data are given to the algorithm in the form of a column matrix. As from validation parameter analysis it was found the optimal number of clusters to be 4, number of exemplar chosen is four for clustering FFS. The distribution of data after AP clustering is shown in Figure 5.5. Different symbols represent the class of the urban street. Four different colours used for four different urban street classes from Figure 5.5

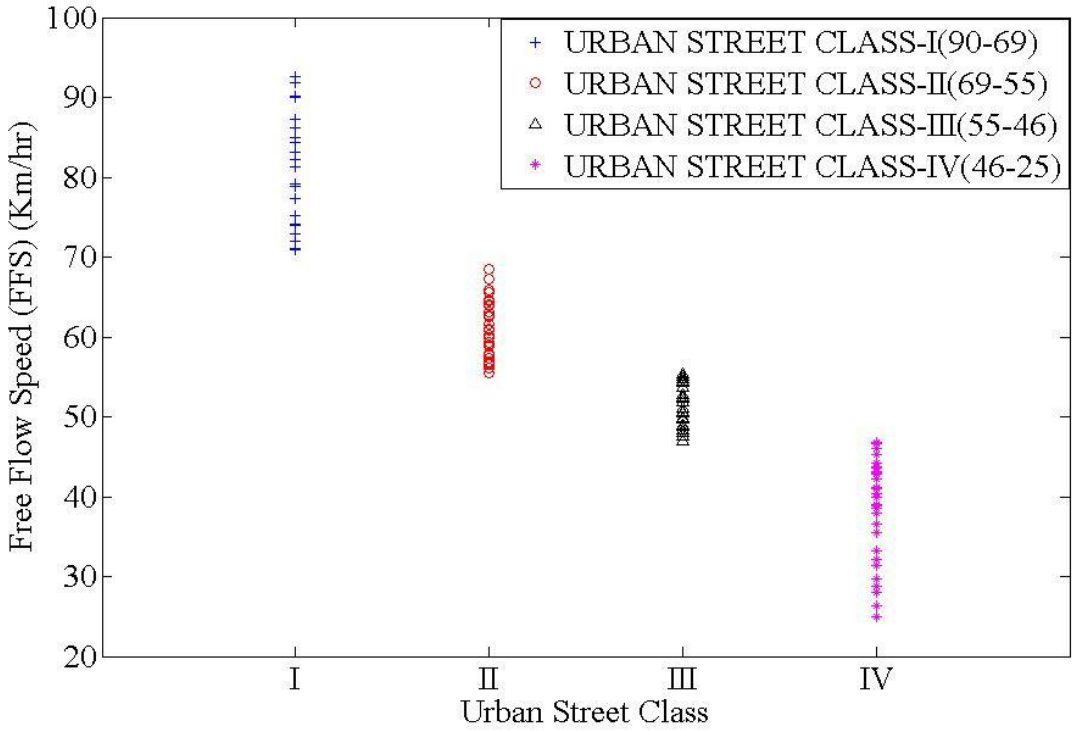
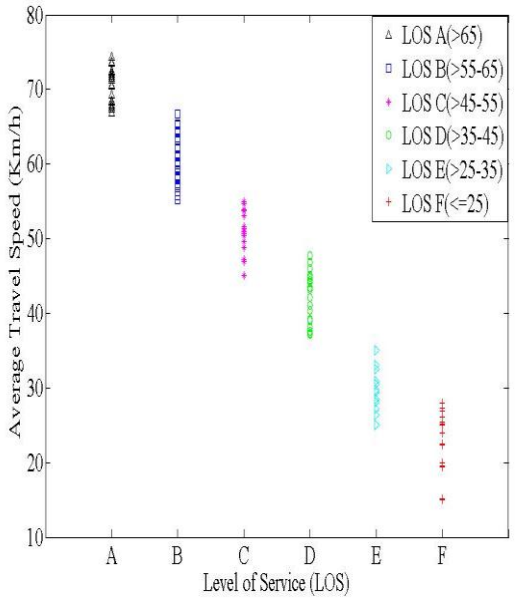
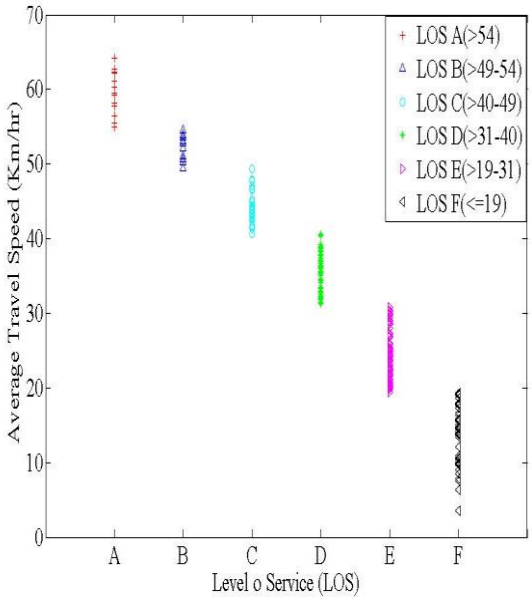


Figure 5.5 SOTA Clustering of FFS for Urban Street Classification

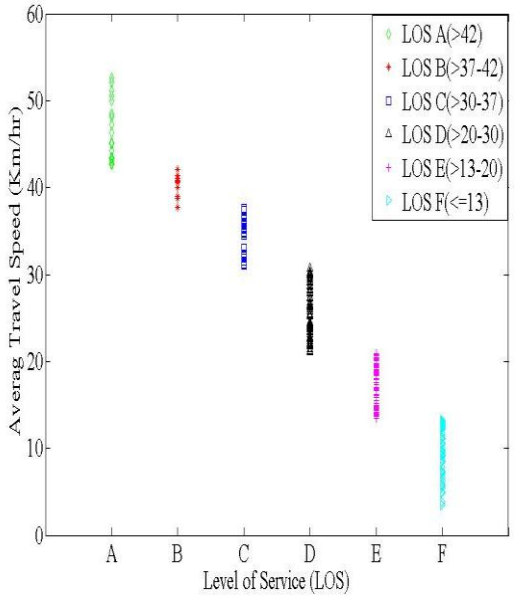
SOTA clustering is used for the second time after classification of urban street segments into four different urban street classes. For four classes of urban street classes LOS speed ranges were defined using SOTA clustering. In this clustering direction wise average travel speed data of both peak and off peak hours were given as input to the clustering algorithm. In Figure 5.6 the clustering result are complied. The legends in Figure 5.6 (A-D) give the speed ranges for the six LOS categories obtained by using SOTA clustering. The speed ranges of LOS categories found using SOTA clustering is also shown in Table 5.2.



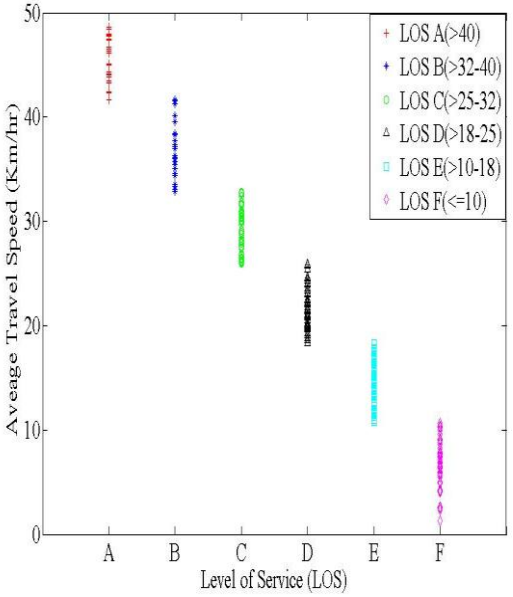
B:LOS of Urban Street Class II



A: LOS of Urban Street Class I



C: LOS of Urban Street Class III



D: LOS of Urban Street Class IV

Figure 5.6 Level of service of urban street classes (I-IV) using SOTA clustering on average travel speed

Table 5.2 Urban Street Speed Ranges for different LOS Proposed in Indian Conditions by SOTA Method

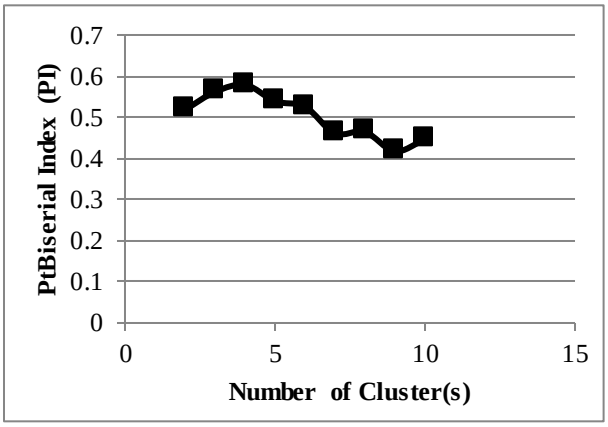
Urban Street Class	I	II	III	IV
Range of free-flow speed (FFS)	90 to 69 km/h	69 to 55 km/h	55 to 46 km/h	46 to 25 km/h
Typical FFS	75km/h	60km/h	47km/h	35 km/h
LOS	Average Travel Speed (Km/h)			
A	>65	>54	>42	>40
B	>55-65	>49-54	>37-42	>32-40
C	>45-55	>40-49	>30-37	>25-32
D	>35-45	>31-40	>20-30	>18-25
E	>25-35	>19-31	>13-20	>10-18
F	≤25	≤19	≤13	≤10

Average travel speed of LOS categories (A-F) expressed in percentage of free flow speeds were found to be approximately 90 and above, 70-90, 55-70, 45-55, 30-45 and less than equal to 30 respectively. Whereas in HCM (2010) it has been mentioned that these values are 85 and above, 67-85, 50-67, 40-50, 30-40 and less than equal to 30 percentage respectively.

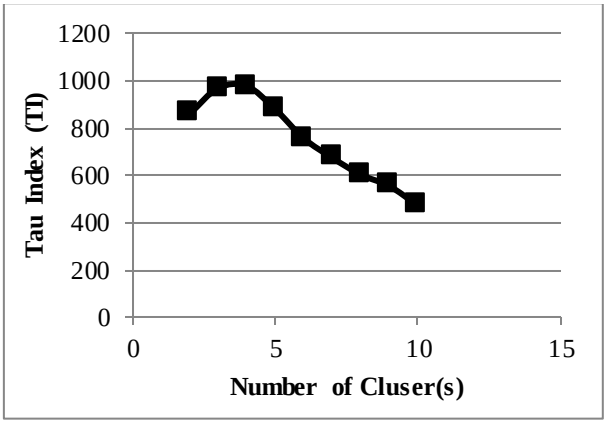
5.2.3 Hard Competitive Learning (*hardcl)* Clustering

Hard Competitive Learning is used for clustering of free flow speed of urban street segments. From this clustering the type of urban class particular segment belongs to is determined than the average travel speed of each segment is used for clustering process to know the speed range for a particular LOS. Like the other two algorithms, free flow speed is used for computing the cluster

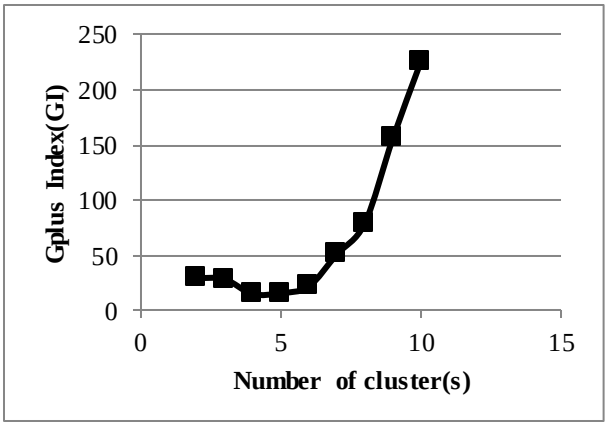
validation parameters. The values of the six validation measures obtained for 3 to 7 number of clusters are plotted from figure 5.7 (A) to 5.7 (F).

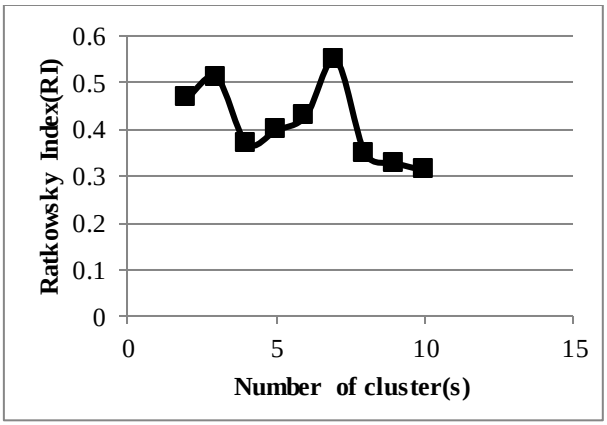


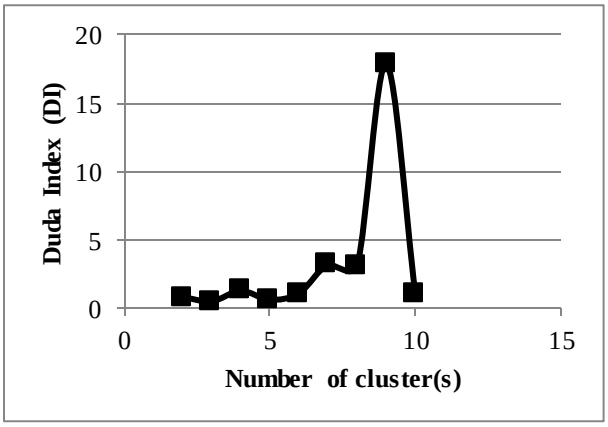
A: PI vs Number of clusters



B: TI vs Number of clusters

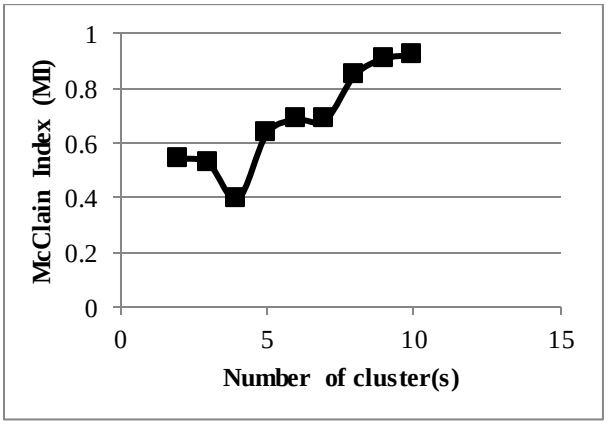






C: GI vs Number of clusters

D: RI vs Number of clusters



E: DI vs Number of clusters

F: MI vs Number of clusters

Figure 5.7 Validation measures for optimal number of clusters using *hardcl* clustering

Six validation parameters are interpreted to obtain the optimum number of clusters for deciding the classification of street segments into different urban street classes. If variation in parametric values from one cluster to the next cluster is not significant it is always considered to go for lesser number of clusters. From Literature review it was believed that the maximum value of the PtBiserial Index (PI) signifies the optimal number of clusters for a particular set of data. The index value is highest for four numbers of clusters which is shown in Figure 5.7 (A). Again the maximum value Tau Index (TI) gives the optimal number of clusters. The Index value is highest for four numbers of clusters as shown in figure 5.7 (B). Also the minimum value of GPlus Index (GI) gives the optimal number of clusters. The index value is minimum for four number of clusters as shown in figure 5.7 (C). The maximum value of Ratkowsky Index (RI) gives the optimal number of clusters. The index value is maximum for seven number of clusters as shown in figure 5.7 (D). Figure 5.7 (E) shows Duda Index (DI). The minimum value of the index gives the optimal number of clusters. So, going with the literature three is taken as the optimal number of cluster. The minimum value of McClain Index (MI) gives the optimal number of clusters. The index value is obtained minimum for four numbers of clusters as shown in figure 5.7 (F). Thus this goes in hand with PI, TI and GI. Hence four out of six validation parameters considered in this study give the optimal cluster value as 4 which is also same as suggested by HCM-2000. That is the reason for which in this research the urban street segments were classified into four Classes by using *hardcl* clustering technique.

In this study data collected from five major urban corridors of Mumbai city comprising of 100 street segments were analyzed. Second wise free flow speed data collected during the night hours using GPS receiver are averaged over each segment. Average of these averaged values taken for

each travel run on street segments are used by a *hardcl* algorithm for the classification of street segments into number of classesThe result obtained using the *hardcl* algorithm for clustering purpose has been illustrated in Figure 5.8. Different types of symbols are used for each type of urban street class.

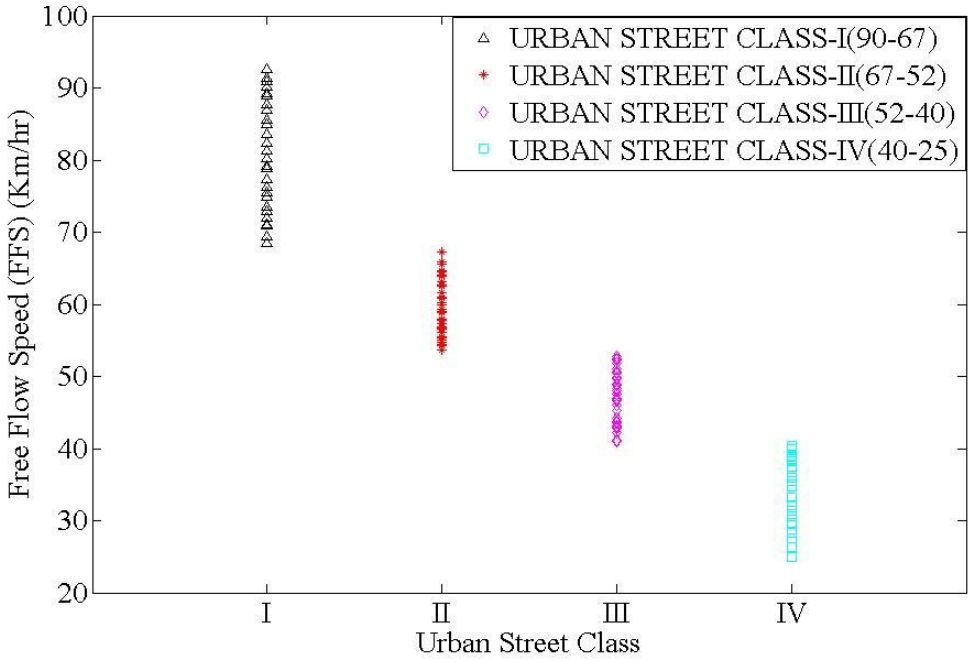
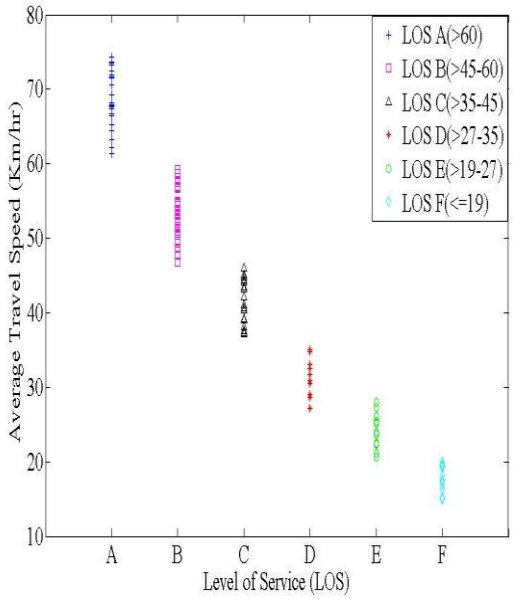


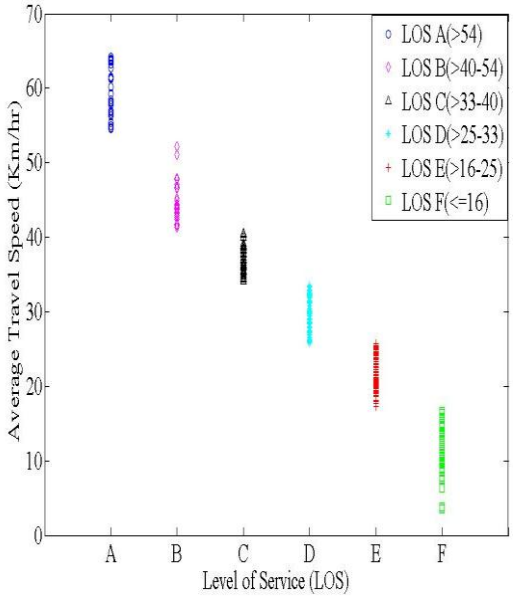
Figure 5.8 *hardcl* Clustering of FFS for Urban Street Classification

The same algorithm was used for second time to the average travel speed data acquired during peak and off peak hour of the above stated five urban street corridors. The *hardcl* algorithm clustered the average speed data into six clusters to give the speed ranges of different urban street classes. The result of the clustering is shown in Figure 5.9 (A) to Figure 5.9(D). Each Level of Service (LOS) for a particular urban street class illustrated in the figure with a unique symbol. The speed ranges for each individual Level of Service categories elaborated in Table 5.3. It can be outlined that the free flow speed range of urban street classes and speed ranges of level of

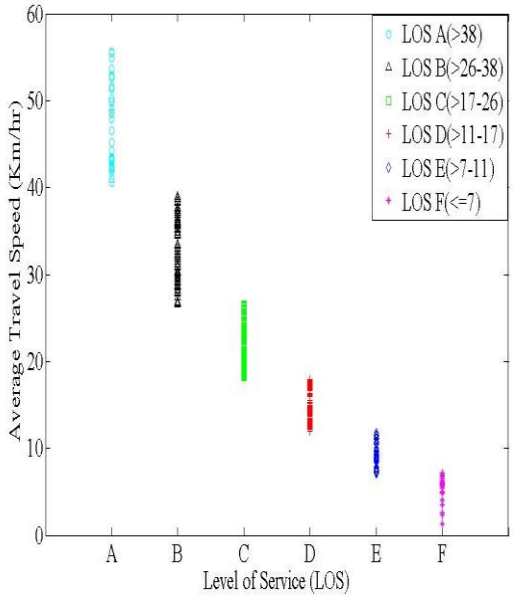
service categories that was resulted from this cluster analysis are significantly lower than that stated in HCM-2000.



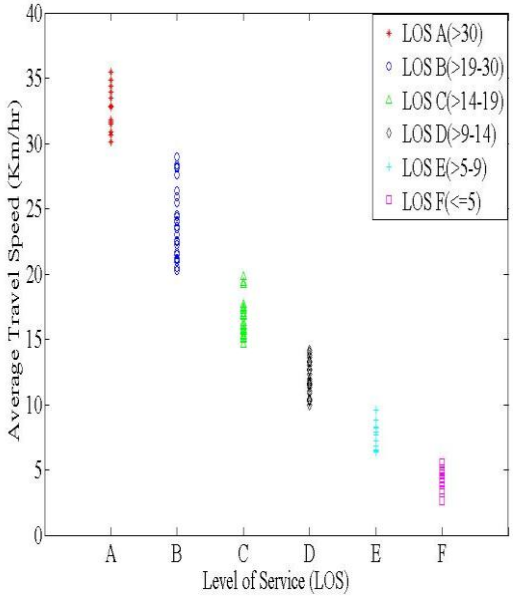
A: LOS of Urban Street Class I



B: LOS of Urban Street Class II



C: LOS of Urban Street Class III



D: LOS of Urban Street Class IV

Figure 5.9 Level of service of urban street classes (I-IV) using hardcl clustering on average travel speeds

Table 5.3 Urban Street Speed Ranges for different LOS Proposed in Indian Conditions by *hardcl* Method

Urban Street Class	I	II	III	IV
Range of Free Flow Speed (FFS)	90 to 67km/h	67 to 52 km/h	52 to 40 km/h	40 to 25km/h
Typical FFS	77km/h	63km/h	50km/h	35 km/h
LOS	Average Travel Speed (Km/h)			
A	>60	>54	>38	>30
B	>45-60	>54-40	>26-38	>19-30
C	>35-45	>33-40	>17-38	>14-19
D	>27-35	>25-33	>11-17	>9-14
E	>19-27	>16-25	>7-11	>5-9
F	≤19	≤16	≤7	≤5

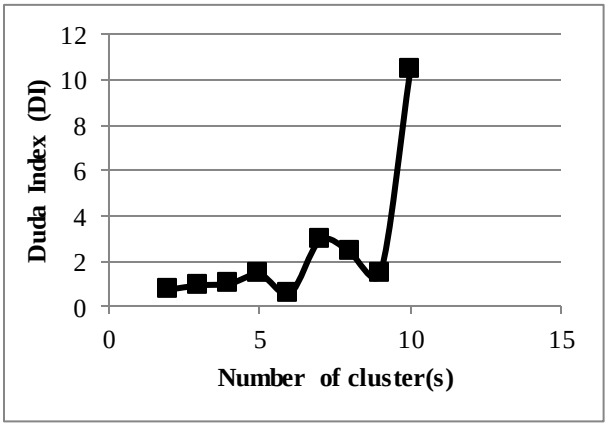
Average travel speed of LOS categories (A-F) expressed in percentage of free flow speeds were found to be approximately 80 and above, 55-80, 35-55, 25-35, 20-35 and less than equal to 20 respectively. Whereas in HCM (2010) it has been mentioned that these values are 85 and above, 67-85, 50-67, 40-50, 30-40 and less than equal to 30 percentage respectively.

5.2.4 Neural Gas (ngas) Clustering

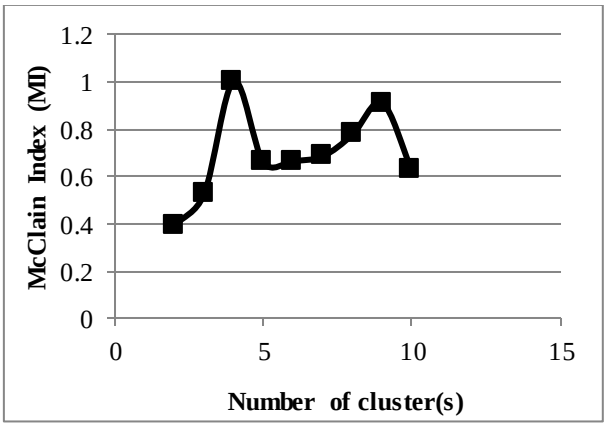
The free flow speed data acquired through GPS receiver was clustered using the ngas algorithm. Neural gas algorithm sorts the network units according to the distance of their reference vector to each input. Then the reference vectors are adapted so that the ones related to the first nodes in the rank order are moved closer than the others to the considered input.

In this research six validation parameters were used. Value of validation parameters were obtained for 4 to 7 numbers of clusters and were plotted in Figure 5.10 (A) to Figure 5.10 (F). These six number of validation parameters were used to know the optimum number of cluster for

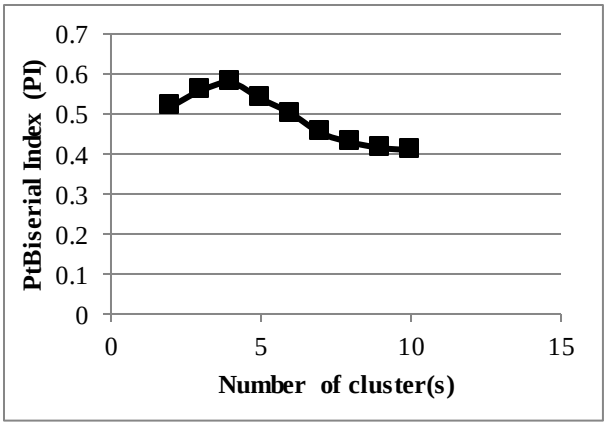
this particular data set of free flow speed. By knowing the optimum number of clusters we can classify the urban street segments into that number of Urban street classes. It is always considered that a lesser number of clusters are better if variation in validation parameters is minimal. According to the available literature the minimum value of Duda Index signifies the optimal number of clusters for a particular data set. Figure 5.10(A) shows that the Index is minimum for 6 number of clusters. The available literature says that the minimum value of Mcclain Index gives the optimal number of clusters. Figure 5.10 (B) shows that the Index are minimum for 4 number of clusters. Figure 5.10(C) shows PtBiserial Index. The maximum value of the index gives the optimal number of clusters. So going with the literature 4 is taken as the optimal number of clusters. For Gplus Index, the smaller value gives the optimal number of clusters. Figure 5.10 (D) shows that the optimal number of clusters are obtained to be 4. The available literature depicts larger values of Tau Index and Ratkowsky Index gives the optimal number of clusters. Figure 5.10 (E), (F) shows that the optimal number of clusters to be 4. Out of six validation parameters considered in this study four parameters give the optimal cluster value as 4 which is also same as suggested by HCM-2000. That is the reason for which in this research the urban street segments were classified into four numbers of classes by using Neural Gasalgorithm.



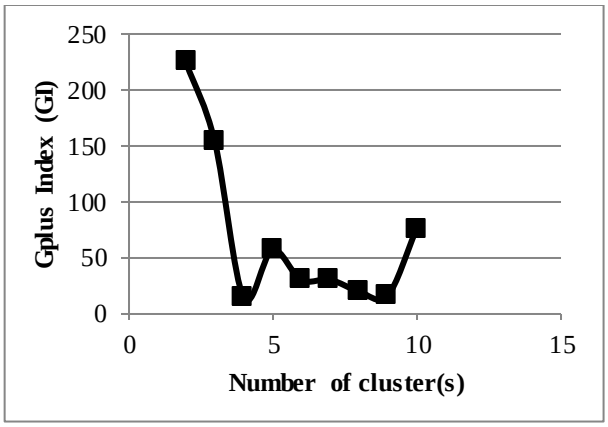
A: DI vs Number of clusters



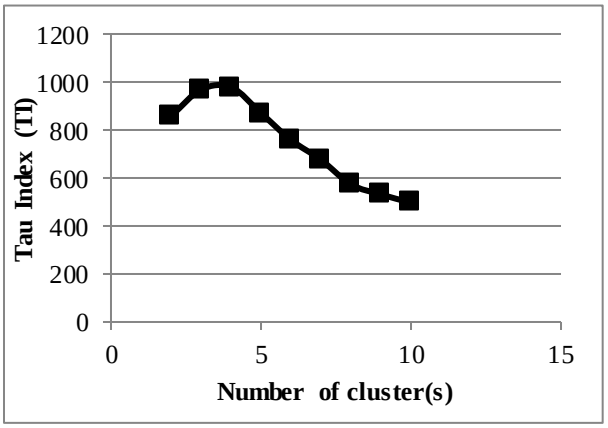
B: MI vs Number of clusters



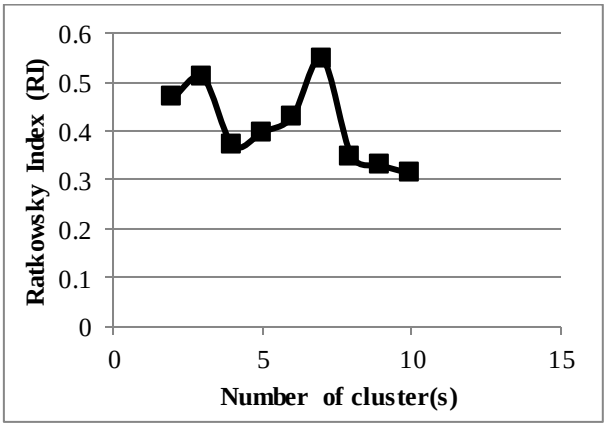
C: PI vs Number of clusters



D: GI vs Number of clusters



E: TI vs Number of clusters



F: RI vs Number of clusters

Figure 5.10 Validation measures for optimal number of clusters using ngas clustering

After determining the optimal number of clusters the FFS data were clustered using ngas algorithm. FFS ranges for each urban street class were found. The clustering result of gas for the urban street class is shown in Figure. 5.11. Various symbols and colors are used to illustrate the speed ranges of various urban street classes.

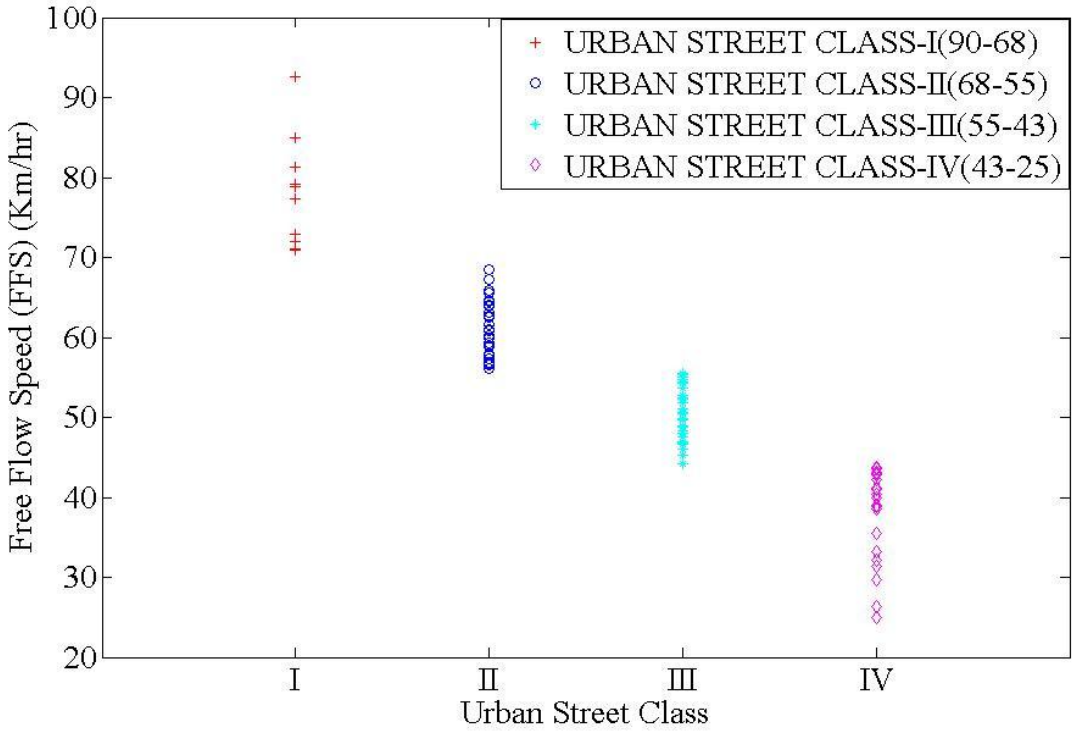
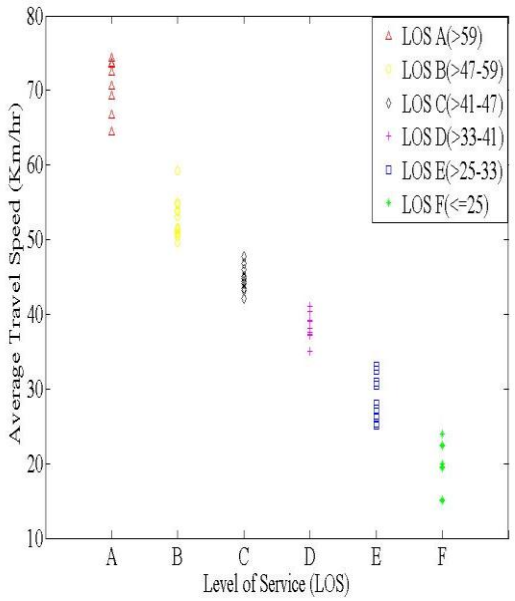
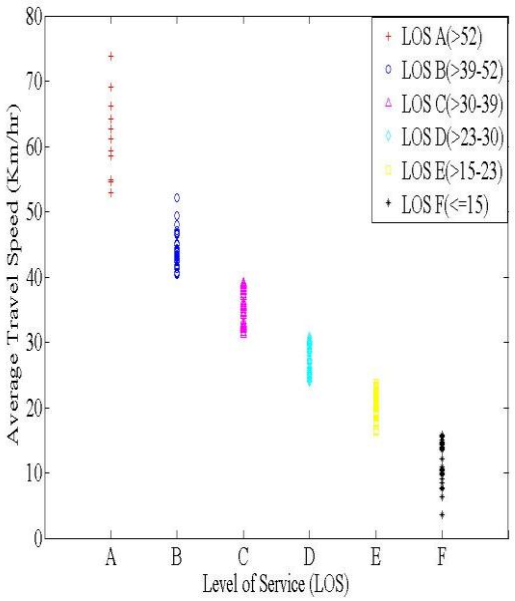


Figure 5.11 Neural Gas Clustering of FFS for Urban Street Classification

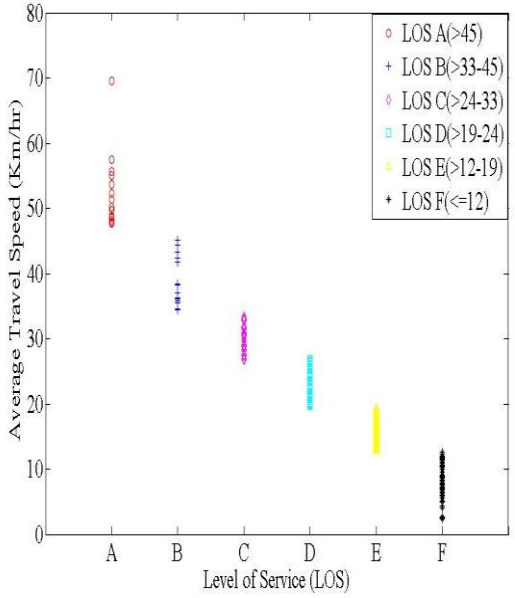
After classification of urban streets into number of classes, average travel speed collected on street segments during both peak and off peak hours were clustered using neural gas to find the speed range of level of service categories. In fig. 5.12 the speed values are shown by different symbols depending on to which LOS category they belong. The legends in fig. 5.12 (A-D) gives the speed ranges for the six LOS categories obtained by using neural gas clustering. The speed ranges for LOS categories found using neural gas clustering is also shown in Table 5.4.



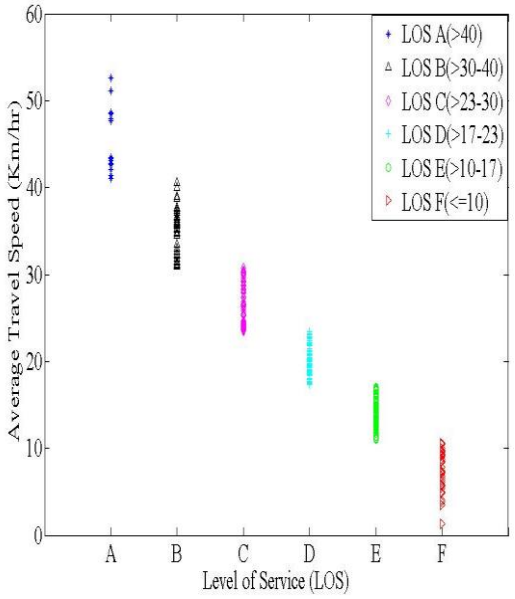
A: LOS of Urban Street Class I



B: LOS of Urban Street Class II



C: LOS of Urban Street Class III



D: LOS of Urban Street Class IV

Figure 5.12 Level of service of urban street classes (I-IV) using ngas clustering on average travel speeds

Table 5.4 Urban Street Speed Ranges for different LOS Proposed in Indian Conditions by ngas Method

Urban Street Class	I	II	III	IV
Range of free-flow speed (FFS)	90 to 68 km/h	68 to 55 km/h	55 to 43 km/h	43 to 25 km/h
Typical FFS	75km/h	60km/h	47km/h	35 km/h
LOS	Average Travel Speed (Km/h)			
A	>59	>52	>45	>40
B	>47-59	>39-52	>33-45	>30-40
C	>41-47	>30-39	>24-33	>23-30
D	>33-41	>23-30	>19-24	>17-23
E	>25-33	>15-23	>12-19	>10-17
F	≤25	≤15	≤12	≤10

Average travel speed of LOS categories (A-F) expressed in percentage of free flow speeds were found to be approximately 80 and above, 60-80, 50-60, 35-50, 30-35 and less than equal to 30 respectively. Whereas in HCM (2010) it has been mentioned that these values are 85 and above, 67-85, 50-67, 40-50, 30-40 and less than equal to 30 percentage respectively.

5.3 Summary

Four clustering algorithms i.e. CLARA, SOTA, hardcl and ngas were used for clustering the data obtained from five corridors of Mumbai city. Several cluster validation measures were used to get the optimal number of clusters. Optimal number of clusters helps us to divide the urban street segments into that number of urban street classes. For this purpose the FFS data were clustered to get the FFS ranges of different urban street classes. The clustering algorithm is used for the second time on the average travel speed data to get the speed ranges of different LOS categories. For every clustering method, the results of analysis are represented graphically as well as in a tabular format.

The next chapter lays emphasis on summary, conclusion, limitation and future work possible.

Chapter 6

Summary, Conclusions and Future Work

6.1 Summary

In this study an attempt has been made to develop methodologies to define LOS criteria for urban streets in Indian context. Cluster analysis has been used in this study for the classification of urban streets and level of service categories. CLARA, SOTA, hardcl and ngas are the four clustering methods used for this study. These clustering methods are used to classify the road segments into various classes and also to define the speed ranges of the LOS. For the collection of these speed data GPS has proven to be a very useful tool and GIS was used to manage these data. The clustering algorithms were used in two stages. In the first stage FFS data were clustered into four different groups corresponding to different urban street classes. In the second stage these clustering algorithms are applied on the average travel speeds to classify into six categories for six levels of service. So, the speed ranges for level of service categories were defined. The optimum number of cluster was determined using various cluster validation parameters. In this study 14 validation parameters were used namely Calinski-Harabasz Index, Connectivity Index, Average Proportion of Non-overlap (APN) Index, Average Distance (AD) Index, Average Distance between Means (ADM) Index, Figure of Merit (FOM) Index, Homogeneity Index (HI), Stability Index (SI), Duda Index, PtBiserial Index, Gplus Index, Tau Index, Ratkowsky Index, McClain Index.

6.2 Conclusion

The following conclusions were derived from the present study in defining Level of Service (LOS) in Indian Context:

- GPS has proved to be a powerful tool in collection of speed data with high accuracy. Large number of data can be collected in a short interval of time. Hence GPS can be used to collect speed data for developing as well as developed countries and cluster analysis can be used to define the speed ranges of LOS categories.

- Four clustering methods i.e. CLARA, SOTA, hardcl and ngas were used for this study purpose. Various cluster validation parameters were used to get the optimal number of clusters. Basing on the optimal number of clusters the urban street was classified into four classes (I-IV). Free flow speed ranges for different urban street classes were found out for each clustering method. The speed ranges were different for different method and lower than HCM (2000). The average travel speed of LOS categories (A-F) expressed in terms of free flow speeds were found out. The results obtained in this study closely resembles the values mentioned in HCM (2010). Roads with varied geometric and surrounding environmental characteristics and heterogeneous traffic flow are responsible for these lower values in FFS.
- The speed ranges of LOS categories were found using the four clustering methods. The speed ranges were found to be different for different clustering method. The speed ranges of different LOS categories are found to be lower than suggested in HCM 2000. The reason behind such lower speeds can be roadside vending, side friction developed due to on-street parking etc.
- The results of this study depict that less number of roads in Mumbai belongs to high speed design (Street Class I) or highly congested (Street Class IV). Majority of the road segments Delong to suburban (Street Class II) or intermediate (Street Class III) type. Hence substantial geometric improvements have to be done in the Greater Mumbai region to mitigate the burden on the urban road infrastructure.

6.3 Limitations and Future Scope

- This study was done using midsized vehicle for the purpose of data collection. The percentage of midsized vehicle is more in urban roads of India and data collected using these vehicles is convenient and easy, but to get a complete scenario of heterogeneous traffic flow further study can be done using more number of modes.
- This study was conducted for the city of Mumbai. Similar study can be carried for other cities of India, with a large diversity of people and their driving characteristics, physical features of the lands and climatic conditions.

- User perception should be considered in defining LOS criteria. Along with quantitative analysis, qualitative analysis should be done and a relation should be developed between quantitative and qualitative study.

References

Arasan, V.T., Vedagiri,P. Study of the impact of exclusive bus lane under highly heterogeneous traffic condition. *Public Transport* ,2010, Vol. 2(1),p. 135-15.

Azimi, M. and Zhang, Y. Categorizing Freeway Flow Conditions by Using Clustering Methods. *Journal of the Transportation Research Board*, 2010, 2173, pp. 105-114.

Baumgartner, W.E. Level of service: Getting ready for the 21st century. *ITE Journal (Institute of Transportation Engineers),* 1996, Vol. 66 (1), p. 36-39

Boomija, M.D. Comparison of partition based clustering algorithms. *Journal of Computer Applications*, 2008, Vol. 1, No. 4 pp.18-21.

Brilon, W. (2000). Traffic flow analysis beyond traditional methods. *Transportation Research Circular E-C018: 4th International Symposium on Highway Capacity*, Transportation Research Board, Washington, D.C., 26–41.

Caliński, R.B., Harabasz, J., A dendrite method for cluster analysis. *Communications in Statistics*, 1974, vol. 3, pp. 1-27.

Camastra, F., Vinciarelli, A. Combining neural gas and learning quantization for cursive character recognition. *Neurocomputing*, 2003, vol. 51, pp. 147-159.

Cameron, R. (1996). G3F7: An expanded LOS gradation system. *ITE Journal*, January, 40-41.

Cheol, O., and Stephen, G. R. (2002). Real-time inductive-signature-based level of service for signalized intersections. *Transportation Research Record*, 1802, Transportation Research Board, Washington, D.C., 97–104.

Clark, I. Level of Service F: Is it really bad as it gets? In Proc.,*IPENZ Transportation Group Conference*, New Plymouth, November, 2008.

Dandan,T., Wei,W., Jian,L., Yang, B. Research on Methods of Assessing Pedestrian Level of Service for Sidewalk., *J Transpn Sys Eng & IT* , 2007, 7(5), pp.74−79.

Datta, S. and Datta, S. Comparisons and validation of statistical clustering tech-niques for microarray gene expression data. *Bioinformatics*, 2003, vol. 19(4), pp. 459-66.

Datta, S. and Datta, S. Methods for evaluating clustering algorithms for gene ex-pression data using a reference set of functional classes. *BMC Bioinformatics*, 2006, vol. 7, pp. 397.

Dimitriadou, E. et al. A quantitative comparison of functional MRI cluster analysis. *Elsevier Science*, 2004, vol. 31(1), pp. 57-71.

Fang, F.C., Pecheux, K.K. Fuzzy data mining approach for quantifying signalized intersection level of services based on user perceptions. *Journal of Transportation Engineering,* ASCE, 2009, 135 (6), pp.349-358.

Flannery, A., Rouphail, N., Reinke, D. Analysis and Modeling of Automobile Users' Perceptions of Quality of Service on Urban Streets. *Transportation Research Record,* 2071, Transportation Research Board, Washington, D.C., 2008, ASCE ,p. 26-34.

Flannery, A., Wochinger, K., and Martin, A. (2005). Driver assessment of service quality on urban streets. *Transportation Research Record,* 1920, Transportation Research Board, Washington, D.C., 25–31.

Handl, J.,Knowles, J., and Kell, D.B. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 2005, vol. 21(15), pp.3201-12.

Herrero, J. et al. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics, 2001, vol. 17(2), pp. 126-36.

Kikuchi,S., Chakroborty,P. Frameworks to Represent the Uncertainty when Determining the Level of Service. *Transportation Research Record,* 1968, 2007, pp. 53-62.

Kita, H., Fujiwara, E. Reconsideration on the level of service and a proposed measure. In Proc., *15th Annual Meeting of JSTE*, Japanese, 1995, p. 25–28.

Kittelson, W.K., Roess, R.P. Highway capacity analysis after the highway capacity manual 2000. *Transportation Research Record*, 1776, Transportation Research Board, Washington, *D.C.* 2001, p. 10–16.

Maitra, B., Sikdar, P.K., Dhingra, S.L. Modelling congestion on urban roads and assessing level of service. *Journal of Transportation Engineering,* 1999, Vol. 125 (6), ASCE, p. 508-514.

Martinetz, T.M. Neural gas network for vector quantization and its application to time series prediction. *IEEE Transactions on Neural Networks*, 1993, vol. 4, pp. 558-569.

Marwah, B.R., Singh, B. Level of service classification for urban heterogeneous traffic: A case study of Kanpur metropolis. In Proc., *Fourth International Symposium on Highway Capacity*, Hawaii, June-July, 2000, p. 271-286.

Mateos, A. et al. Supervised neural network for clustering conditions in DNA array data after reducing noise by clustering gene expression profiles. *Methods of Microarray Data analysis II*, 2002, pp. 91-103.

Milligan, G., Cooper, M. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 1985, vol. 50(2), pp. 159-179.

Murugavel, P. and Punithavalli, M. Improved hybrid clustering and distance based technique for outlier removal. *International Journal on Computer Science and Engineering*, 2011, Vol. 3, No. 1, pp.333-339.

Ndoh N.N. Ashford N.J. Evaluation of transportation level of service using fuzzy sets. 73rdAnnual Meeting of TRB, *Transportation Research Board*, Washington, DC, 1994.

Pattnaik S.B., Kumar K.R. Level of service of urban roads based on users perception. *Civil Engineering Systems*, 1996, Vol. 14, p. 87-110.

Pecheux, K., Flannery, A., Wochinger, K., Lappin, J., and Rephlo, J. (2004). Automobile drivers' perceptions of service quality on urban streets. *Transportation Research Record,* 1883, Transportation Research Board, Washington, D.C., 167-175.

Prassas, E.S., Roess, R.P., and Mcshane, W.R (1996). Cluster analysis as tool in traffic engineering. *Transportation Research Record,* 1551, TRB, National Research Council, Washington, D.C., 39–48.

Ratkowsky, D.A., Lance, G.N. A criterion for determining the number of groups in a classification. *Australian Computer Journal*, 1978, vol. 10, pp.115-117.

Rholf , F. Methods of comparing classifications. *Annual Review of Ecology and Systematics*, 1974, vol. 5, pp. 101-113.

Shao, M., Sun, L. United evaluation model of traffic operation level for different types of urban road. Journal *of Tongji University*, 2010, Vol. 38 (11) ,p. 1593-1598.

Spring, G. S. Integration of safety and the highway capacity manual. In *Proc., 4th International Symposium on Highway Capacity, Transportation Research Board, Washington, D.C*., 1999, p. 63–72.

Taylor, M.A.P., Woolley, J. E., and Zito, R. Integration of the global positioning system and geographical systems for traffic congestion studies. *Transportation Research Part-C*, 2000, Vol. 8, 257–285.

Turner,S.M., Eisele,W.L., Benz, R.J., Holdener, D. J. *Travel time data collection handbook*, Texas Transportation Institute, The Texas A&M Univ. System, College Station, Texas, 1998.

Vascak, J. Using neural gas networks in traffic navigation. Acta Technica Jaurinensis, 2009, vol. 2(2), pp. 203-215.

Wang, H. et al. Self-Organizing tree –growing network for the classification of protein sequences. *Protein Science*, 1998, vol. 7, pp. 2613-2622.

Wei, C.H., Chang, C.C., and Wang, S.S (1996). Vehicle classification using advanced technologies. *Transportation Research Record*, 1551, TRB, National Research Council, Washington, D.C., 45–50.

Yang, H., and Qiao, F. (1998). Neural network approach to classification of traffic flow states. *Journal of Transportation Engineering*, ASCE, 124(6), 521–525.

Yang, M. S. et al. On tree types of competitive learning algorithms with their comparisons and applications of MRI segmentation. *International Journal of Intelligent Systems*, 2010, vol. 25(11), pp. 1081-1102.

Appendix-I

The table illustrates FFS and Average travel speed data of 13 segments belonging to Corridor-1. Both these data collection procedure are elaborated in Chapter-3.

Table AI. FFS and Average travel speed during peak ans off peak hours of Corridor-1

Corridor-1					
Segment No.	Average Free-Flow Speed (Km/hr)	Duration and Direction of Travel			
		M-N-S	M-S-N	E-N-S	E-S-N
		Average Travel Speed (Km/hr)			
1	85.00	73.38	73.63	73.53	43.20
2	92.56	68.64	49.56	59.25	38.08
3	72.85	54.65	55.01	39.25	54.66
4	50.66	29.97	38.16	36.355	31.25
5	39.89	27.36	21.60	26.465	18.43
6	38.58	31.82	34.94	26.39	25.73
7	53.61	20.88	19.51	24.745	21.53
8	48.00	18.78	23.84	25.795	18.14
9	46.92	12.01	21.29	7.285	14.85
10	49.64	18.68	17.60	24.915	15.48
11	38.91	12.80	8.29	21.02	12.77
12	40.39	17.71	22.19	28.06	22.81
13	41.19	38.61	24.00	31.57	19.81

Note:

M-N-S= Morning-North-South

M-S-N=Morning-South-North

E-N-S= Evening-North-South

E-S-N= Evening-South-North

List of Publications

Journals:

1. Das, A.K., Bhuyan, P.K. Level of Service Criteria of Urban Streets using Clustering Large Applications (CLARA). *Advances in Transportation Studies*. (Submitted).
2. Patnaik, A.K., Das, A.K, Dehury, A.N., Bhuyan, P.K., Chattaraj, U., Panda, M., Adaboost clustering in defining LOS criteria of Mumbai city. *IJEI-e-ISSN:-2278-7461, P-ISSN:-2319-6491*, 2013, Volume 2, Issue-8, pp. 45-55.
3. Das, A.K., Patnaik, A.K., Dehury, A.N., Bhuyan, P.K., Chattaraj, U., Panda, M., Defining Level of Service criteria of urban streets using neural gas clustering. 2013, *IOSRJEN, (Accepted)*.
4. Dehury, A.N., Patnaik, A.K., Das, A.K., Chattaraj, U., Bhuyan, P.K., Panda, M., Accident analysis and modelling on NH-55. *IJEI, ISSN: 2278-7461, ISBN-2319-6491, 2013*, Volume 2, Issue-7, pp. 80-85.
5. Dehury, A.N., Patnaik, A.K., Das, A.K., Chattaraj, U., Bhuyan, P.K., Panda, M., Black spot analysis on national highways. *IJERA, ISSN:-2248-9622, 2013*, vol.(3), issue 3,pp. 402-418.
6. Dehury, A.N., Patnaik, A.K., Das, A.K., Chattaraj, U., Bhuyan, P.K., Panda, M., Accident analysis on two lane highways. *IJETAE, ISSN:-2250-2459*, 2013, Volume 2,issue-6. *(Submitted)*.
7. Das, A.K., Bhuyan, P.K., Defining LOS criteria of urban streets using SOTA clustering. *Periodica Polytechnica.(Under preparation)*